Essays in Microeconomic Theory

Thesis advisor

Professor Jerry R. Green

Author

Lauren Nicole Merrill

Essays in Microeconomic Theory

Abstract

If the number of individuals is odd, Campbell and Kelly (2003) show that majority rule is the only non–dictatorial strategy–proof social choice rule on the domain of linear orders that admit a Condorcet winner, an alternative that is preferred to every other by a majority of individuals in pairwise majority voting. This paper shows that the claim is false when the number of individuals is even, and provides a characterization of non–dictatorial strategy–proof social choice rules on this domain. Two examples illustrate the primary reason that the result does not translate to the even case: when the number of individuals is even, no single individual can change her reported preference ordering in a manner that changes the Condorcet winner while remaining within the preference domain. Introducing two new definitions to account for this partitioning of the preference domain, the chapter concludes with a counterpart to the characterization of Campbell and Kelly (2003) for the even case.

Adapting the models of Laibson (1994) and O'Donogue and Rabin (2001), a learning–naïve agent is presented who is endowed with beliefs about the value of the quasi–hyperbolic discount factor that enters into the utility calculations of her future–selves. Facing an infinite–horizon decision problem in which the payoff to a particular action varies stochastically, the agent updates her beliefs over time. Con-

ditions are given under which the behavior of a learning–naïve agent is eventually indistinguishable from that of a sophisticated agent, contributing to the efforts of Ali (2011) to justify the use of sophistication as a modeling assumption.

Building upon the literature on one–to–one matching pioneered by Gale and Shapley (1962), this paper introduces a social network to the standard marriage model, embodying informational limitations of the agents. Motivated by the restrictive nature of stability in large markets, two new network–stability concepts are introduced that reflect informational limitations; in particular, two agents cannot form a blocking pair if they are not acquainted. Following Roth and Sotomayor (1990), key properties of the sets of network–stable matchings are derived, and concludes by introducing a network–formation game whose set of complete–information Nash equilibria correspond to the set of stable matchings.

Contents

List of Figures

Citations to Previously Published Work

Portions of Chapter 2 appear in

> Merrill, L. N. (2001): "Parity Dependence of a Majority Rule Characterization on the Condorcet Domain," *Economics Letters*, 112, 259–261.

Acknowledgments

It is a pleasure to thank those who made this dissertation possible by supporting me, my research, and my professional and personal development during my tenure as a graduate student. I am exceptionally grateful...

...To Professor Jerry Green, who undertook to act as my supervisor and mentor despite his many other academic and professional commitments. His wisdom, knowledge, and commitment to the highest standards inspired and motived me, and it was a singular honor to teach under his guidance.

...To Professor David Laibson and Professor Tomasz Strzalecki, who generously gave of their time and expertise to better my work. I thank them for their contribution and good–natured support. And to Professor Drew Fudenberg, for his unrivaled enthusiasm in teaching me game theory and encouraging my research efforts.

...To Brenda Piquet, without whom all would be chaos.

...To Professor Donald Campbell of the College of William & Mary, for sharing with me his passion for creating knowledge and for considering me his colleague well before the honor was earned. Without his mentorship I might never have have understood the thrill of discovery.

...To Michael Sinkinson, Trisha Banerjee, Serge Ryappo, Sam Espahbodi, and Frederik Denef, for their love and profound understanding. Throughout my graduate career, they kept my spirits high and helped me see the beauty in the mundane.

...To Yuhta Ishii and José Luis Montiel Olea, for their friendship and kindness. Their companionship and unflinching support of my every ambition brightened my days, and it was a pleasure to share office space with them.

...To my students, for providing some of the most rewarding moments in my graduate student career. Through teaching, I discovered much about myself.

...To the Harvard University Department of Economics and the National Science Foundation, for providing me with the financial support that allowed me to focus on my education, my teaching, and my research.

...To my parents, Christine and Kevin Merrill; my brother, Alex Merrill; my grandparents, Beth and Michael Merrill and Martha and Robert Webb. Their unconditional love and support set me on a path to academic excellence that few have the privilege to obtain, and for that I am forever thankful.

...And to my partner, Ashwin Rastogi, without whom this effort would have been overwhelming and without joy. His tenderness, insightfulness, and empathy continually astound me, and words cannot convey the privilege I feel in being allowed to know him and to love him.

Dedicated to my best friend and partner, Ashwin Rastogi,

and to my parents, Christine and Kevin Merrill.

Chapter 1

Introduction

On a fundamental level, the study of microeconomic theory can be understood to be a study in incentives. At their most basic element, economic models illustrate a decision maker who, faced with a sequence of known or expected payoffs that are conditional on an implemented action, must choose from some pre–specified choice or action set so as to maximize her expected payoff. In this context, the structure of the payoff implicitly constitute incentives: the nature of the payoff scheme and the differences in expected payoffs conditional on implementing different actions or making different choices is of sole significance in determining the action that will be implemented by the agent.

Within this work, incentives will be considered through three distinct filters in three divergent modeling environments. Although formally residing in different subfields of microeconomic theory – social choice theory, behavioral economics, and market design – the models of the subsequent three chapters share the common dialogue of incentives.

Focusing first on a model in which incentive questions are relevant at an individual level within a larger social setting, Chapter 2 characterizes non–dictatorial voting rules that satisfy the particular incentive compatibility requirement of strategy–proofness: a voting rule is strategy–proof if it is in the best interest of each individual to vote truthfully (that is, to reveal her true preferences) regardless of the behavior or preferences of other voters. The study of voting rules is only meaningful within the context of a larger society, as voting becomes trivial in a society of a single individual. Nonetheless, the incentive concept is one that focuses on individual behavior by asking that each agent have the unique dominant strategy of voting truthfully.

Chapter 2 focuses on a particular question in social choice theory, and provides a counter–point to an existing majority rule characterization result. Given a set of individuals, N, and a set of feasible alternatives, X, a social choice rule selects a single alternative from X as a function of individual preferences. May (1952) characterized majority rule as the unique social welfare function satisfying Independence of Irrelevant Alternatives, Neutrality, Anonymity, and a strong positive responsiveness axiom, the last of which Maskin (1995) replaced with the Pareto criterion in showing that any social welfare function other than majority rule satisfying the four axioms will fail to be transitive–valued at some preference profile at which majority rule is transitive–valued. Using a weaker set of axioms, Campbell and Kelly (2000) established the same result.

Campbell and Kelly (2003) provide a strategy–proofness characterization of majority rule as the unique non–dictatorial strategy–proof social choice rule when there is an odd number of individuals with strict preferences in the Condorcet domain, the

domain of linear preference profiles at which there exists an alternative that is preferred to every other alternative by some strict majority of individuals. Furthermore, they show that when there is an even number of individuals, if a social choice rule is non–dictatorial and cannot be manipulated by individuals or by two–individual coalitions, then it is majority rule.

In Chapter 2, two examples are given of non–dictatorial strategy–proof social choice rules on the Condorcet domain with an even number of individuals that are distinct from majority rule. The construction of these examples relies on a key feature of the preference domain that is parity–dependent: when there is an even number of individuals, the Condorcet domain can be partitioned into components such that no individual can change her reported preference ordering in a manner that moves the reported preference profile from one partition component to another.

To formalize these structural differences, three lemmas are introduced, which show respectively the existence of a natural partition of the Condorcet domain, the implications of this partition for strategy–proofness, and how strategy–proof rules on the Condorcet domain can be deconstructed into strategy–proof rules on smaller domains via the partition components. Theorem 2.4.4 employs these lemmas in characterizing non–dictatorial strategy–proof rules on the Condorcet domain with an even number of individuals, and Theorem 2.4.11 provides a counterpart to the Campbell and Kelly (2003) result when n is even.

The Condorcet domain is an admittedly restrictive domain; in particular, it is worth noting that an individual considering a manipulation is constrained in her admissible reported preferences by the reported preferences of the other individuals.

Campbell and Kelly (2003) justify consideration of this domain as providing a complement to the Maskin (1995) social welfare function characterization of majority rule. In the context of Chapter 2, consideration of the Condorcet domain is employed to illustrate the crucial dependence of the Campbell and Kelly (2003) characterization of majority rule on the parity of the set of individuals.

The main contribution of this chapter therefore lies in identifying this dependence and illustrating through Examples 2.3.1 and 2.3.2 that the characterization does not hold for an even number of individuals. The axiomatic characterization of non–dictatorial strategy–proof social choice rules on this domain should be viewed as a supplementary result meant to formalize the crucial role that parity plays in the characterization result: when the set of individuals is odd, majority rule is uniquely non–dictatorial and strategy–proof; when the set of individuals is even, there many rules satisfying these criteria, of which majority rule is one of the less exotic members.

Moving into the area of behavioral economics, Chapter 3 constructs a model in which internal incentives are of crucial importance. Here, there is no society to speak of, only an individual agent facing an infinite horizon decision problem while partially aware of her time–inconsistent preferences. Within this single–agent model, incentives interplay between the various contemporaneous selves acting in each period; the agent's behavior today must be a best response to the anticipated actions of her future selves, and may in fact have a large influence on their action sets. In an environment in which the tendency to procrastinate abounds, the classic behavioral types exhibit very distinct equilibrium behavior: the naïve agent, unaware of the time–inconsistency of her preferences, postpones indefinitely the completion of a task even

when delay is costly; the sophisticated agent, fully aware of her time–inconsistency, completes the task immediately along the equilibrium path.

Chapter 3 introduces a new type of behavioral agent, a learning–naïve agent, whose observed equilibrium behavior in an infinite–horizon decision problem bridges the gap between the observed behavior of naïve and sophisticated agents. Behavioral economics has large focused on two main types of agents with time–inconsistent preferences: naifs and sophisticates. Both types of agents are modeled as a sequence of autonomous temporal selves indexed by time, with per period utility functions characterized by a (β, δ) pair, where $\beta \in (0, 1)$ represents the agent's discounting of all future periods relative to the current period and $\delta \in (0, 1)$ represents the agent's discounting between consecutive periods. Naïve agents are unaware of the dynamic inconsistency of their preferences, and believe that future–period selves will make utility maximizing decisions that coincide with the decision that the current period self would make for them; that is, they believe that future–period will behave as if $\beta = 1$. Hence, utility maximizing naïve agents do not choose to employ costly commitment mechanisms that restrict the behavior of future–period selves. Sophisticated agents, on the other hand, recognize that their preferences are dynamically inconsistent and that without the implementation of commitment mechanisms, future–period selves will make decisions that are suboptimal with regard to their current period utility function.

In infinite–horizon models of costly procrastination, the standard result is for naïve agents to procrastinate indefinitely and for sophisticated agents to employ a commitment mechanism in which the task is accomplished in the first period. Della Vigna

and Malmendier (2006) offer data on healthy club memberships that do not appear to support these results. In particular, note that "low–attendance consumers delaying canceling [their membership]...despite small transaction costs." That is, agents for whom the maintenance of a health club membership is costly relative to their benefits from exercise (i.e., those who do not frequently use the facilities) maintain their membership for a significant number of periods. However, even these agents eventually terminate their membership; in their data on monthly membership uses, over 75% of low–frequency members terminated their membership within twenty–four periods; for comparison, roughly one third terminated membership within 12 periods, and over 60% terminated their membership within 18 periods. As Della Vigna and Malmendier Della Vigna and Malmendier (2006) note, "observed consumer behavior is difficult to reconcile with standard preferences and beliefs."

Where Della Vigna and Malmendier (2006) propose a model of overconfidence in future self–control, Chapter 3 proposes a model of learning in which a learning–naïve agent slowly moves from naïve–type optimization behavior to sophisticated–type optimization behavior by experiencing stochastic per period shocks to her utility function. The learning–naïve agent is initially uncertain about the value of β that enters into the utility functions of her past– and future–period selves, but holds prior beliefs about the value, $\hat{\beta}$. By observing the stochastic shocks and action choices of past selves, the learning–naïve agent is able to update her beliefs about the value of $\hat{\beta}$, the quasi–hyperbolic discount factor of her past– and future–period selves. Given sufficiently many periods and sufficient variation in the stochastic shocks, the equilibrium behavior of the learning–naïve agent eventually converges with that of the sophisti-

cated agent: when the parameters of the model are such that the sophisticated agent would employ a commitment mechanism, the learning–naïve agent will eventually do the same.

Returning to a non–degenerate society of many individuals, Chapter 4 adapts the standard one–to–one matching literature to an environment in which individual agents have very limited information about the market in which they interact. In particular, agents are assumed to only be acquainted with a certain subset of the population. Here, the equilibrium concept rests on pair–compatible incentives: in the proposed marriage market with local information, a matching is stable if there exist no pairs of individuals who are acquainted with one another and who prefer to be matched to one another than to their partners has prescribed by the matching. Here, verifying the incentive compatibility of a proposed match requires joint consideration of two individuals preferences simultaneously.

Chapter 4 presents a model of one–to–one matching in which agents have limited information about the composition of the matching market. In the standard one–to–one matching model of Gale and Shapley (1962), agents are implicitly assumed to have complete knowledge of the identities of the other agents in the market, their own preferences over these agents, and the preferences of the other agents. In Chapter 4, the standard one–to–one two–sided matching model is augmented by introducing a concept of local stability. Notions of distance are encoded in a social network, in which directly linked agents are viewed as being "close" to one another. This network structure allows for the introduction of two new definitions of stable matchings, depending on the interpretation of the network. Under the first definition, the net-

work is seem as limiting the set of potential blocking pairs: only agents who are linked in the network share enough information about one another's preferences to block a potential match. Under the second definition, the network is seen as additionally imposing direct restrictions on the pairs of agents who can be matched: agents who are not located sufficiently close enough – that is, who are not linked – cannot be matched to one another at any stable matching.

Following Roth and Sotomayor (1990), the analogous properties of these two sets of network–stable matchings are derived for comparison to the standard model. Through the implementation of an augmented Deferred Acceptance Algorithm (Gale and Shapley, 1962), the existence of both types of network–stable matching is shown for generic marriage networks. Moreover, nesting relations between the two new network–stable matching sets and the set of stable matchings on the associated marriage network are derived, and examples are provided to illustrate cases in which nesting does not generally hold. Finally, the marriage network framework is imagined to be preceded by a network formation game, in which agents simultaneously propose links while only have access to limited information about the market.

The model presented here differs in many crucial respects from the literature on strategic network formation. Jackson and Wolinsky (1996) model a dynamic setting in which the network structure evolves as agents reoptimize their sets of connections to maximize utility arising from communication across the network, and show that efficient networks need not be stable in such an environment. Using an implementation approach, Dutta and Mutuswami (1997) show that this tension can be partially reconciled in certain settings. Bala and Goyal (1999) characterize of the architecture

of equilibrium networks in a setting in which agents receive direct benefits from connecting to other agents in a network, and receive indirect benefits from neighbors' neighbors. Analyzing a similar model, Jackson and Rogers (2005) show that the equilibrium network structures exhibit a certain "small worlds" property, in which densely connected groups of individuals are connected to other groups by sparse links. Bridging the gap between the literatures on network formation games and bargaining on networks, Bloch and Jackson (2007) consider a model with transferable utility in which agents propose and maintain links to maximize direct and indirect benefits received from their network connections. In contrast to these bodies of work, the model of marriage networks presented herein presents the network formation game as preceding a formal matching process, and as such provides microeconomic foundations for utility derived solely from network connections. Within the marriage networks framework, indirect benefits to connections are received insofar as they impact the set of network–stable matchings.

The marriage networks model presented here shares many common properties with that of Arcaute and Vassilvitskii (2009). Whereas this paper focuses on the set theoretic properties of the sets of network–stable matchings, their work provides a reinterpretation of the Deferred Acceptance Algorithm of Gale and Shapley (1962) as a myopic best response dynamic. While providing a similar nesting result on one set of network–stable matchings and stable matchings to the associated marriage problem, Arcaute and Vassilvitskii (2009) approach the question from the perspective of computer science and show that proving the equality of these two sets is NP–complete.

Chapter 2

Parity Dependence of a Majority Rule Characterization

2.1 Introduction

Given a set of individuals, N, and a set of feasible alternatives, X, a *social choice rule* selects a single alternative from X as a function of individual preferences. May (1952) characterized majority rule as the unique social welfare function satisfying Independence of Irrelevant Alternatives, Neutrality, Anonymity, and a strong positive responsiveness axiom, the last of which Maskin (1995) replaced with the Pareto criterion in showing that any social welfare function other than majority rule satisfying the four axioms will fail to be transitive–valued at some preference profile at which majority rule is transitive–valued. Using a weaker set of axioms, Campbell and Kelly (2000) established the same result.

Campbell and Kelly (2003) provide a strategy–proofness characterization of ma-

jority rule as the unique non–dictatorial strategy–proof social choice rule when there is an odd number of individuals with strict preferences in the Condorcet domain. Furthermore, they show that when there is an even number of individuals, if a social choice rule is non–dictatorial and cannot be manipulated by individuals or by two–individual coalitions, then it is majority rule.

This paper provides two examples of non–dictatorial strategy–proof social choice rules on the Condorcet domain with an even number of individuals that are distinct from majority rule. The construction of these examples relies on a key feature of the preference domain that is parity–dependent: when there is an even number of individuals, the Condorcet domain can be partitioned into components such that no individual can change her reported preference ordering in a manner that moves the reported preference profile from one partition component to another.

To formalize these structural differences, three lemmas are presented that show respectively the existence of a natural partition of the Condorcet domain, the implications of this partition for strategy–proofness, and how strategy–proof rules on the Condorcet domain can be deconstructed into strategy–proof rules on smaller domains via the partition components. Theorem 2.4.4 employs these lemmas in characterizing non–dictatorial strategy–proof rules on the Condorcet domain with an even number of individuals, and Theorem 2.4.11 provides a counterpart to the Campbell and Kelly (2003) result when n is even. All proofs have been relegated to the appendix.

The Condorcet domain is an admittedly restrictive domain; in particular, it is worth noting that an individual considering a manipulation is constrained in her admissible reported preferences by the reported preferences of the other individuals.

Campbell and Kelly (2003) justify consideration of this domain as providing a complement to the Maskin (1995) social welfare function characterization of majority rule. In the context of this paper, consideration of the Condorcet domain is employed to illustrate the crucial dependence of the Campbell and Kelly (2003) characterization of majority rule on the parity of the set of individuals.

The main contribution of this note therefore lies in identifying this dependence and illustrating through Examples 2.3.1 and 2.3.2 that the characterization does not hold for an even number of individuals. The axiomatic characterization of non–dictatorial strategy–proof social choice rules on this domain should be viewed as a supplementary result meant to formalize the crucial role that parity plays in the characterization result: when the set of individuals is odd, majority rule is uniquely non–dictatorial and strategy–proof; when the set of individuals is even, there many rules satisfying these criteria, of which majority rule is one of the less exotic members.

Section 2.2 formally introduces the notation and terminology. Section 2.3 contains the main contributions of this note, in the form of two examples illustrating the infeasibility of adapting the Campbell and Kelly (2003) result to the case of an even number of individuals. Section 2.4 formalizes the intuition developed by the examples, and presents a result that is analogous to the Campbell and Kelly (2003) characterization by defining a new social choice rule, *quasi–majority rule*.

2.2 Notation & Definitions

Let $N = \{1, \ldots, n\}$ be a set of individuals and $X = \{x, y, z, \ldots\}$ a set of feasible alternatives, with $\#X \equiv m \geq 3$. Let $A(X)$ denote the set of all complete asymmetric

binary relations on X, and $L(X) \subset A(X)$ the subset of transitive orderings with a maximal element.

Given a *preference profile* $p \in L(X)^n \equiv \mathcal{L}$, let $p(i)$ denote the preference ordering assigned to $i \in N$ by p and let $x \succ_i^p y$ indicate that $i \in N$ (strictly) prefers alternative $x \in X$ to $y \in X$ at profile p. Furthermore, let $p_k(i)$ denote the k-th ranked alternative in $p(i)$, with $p_1(i)$ denoting individual i's most–preferred alternative at p.

A *social choice rule* is a function $g : \wp \to X$ that selects an alternative in X for every preference profile $p \in \wp$. Let $g|_P$ denote the *restriction of* g *to* $P \subset \wp$, where P is a *subdomain* of $\wp$.

A social choice rule $g : \wp \to X$ is *dictatorial* if there exists $i \in N$ such that $g(p) = p_1(i)$ for all $p \in \wp$, and is *unanimous* if $g(p) = x$ whenever $p_1(i) = x$ for all $i \in N$. Furthermore, g is *manipulable* if there exist $p, q \in \wp$ and $i \in N$ such that $p(j) = q(j)$ for all $j \in N \setminus \{i\}$ and $g(q) \succ_i^p g(p)$; that is, if $i \in N$ can *manipulate* g *at* p *via* $q(i)$ by reporting $q(i)$. A social choice rule is *strategy–proof* if it is not manipulable.

2.2.1 The Condorcet Domain

The results of this paper pertain to a particular set of preference profiles known as the Condorcet domain. Alternative x is the *strong Condorcet winner* at profile $p \in A(X)^n$ if

$$\#\{i \in N : x \succ_i^p y\} > \frac{n}{2} \tag{2.1}$$

for all $y \in X \setminus \{x\}$; it is clear that if a strong Condorcet winner exists, it must be unique.

Let $\wp_C \subset A(X)^n$ denote the set of all profiles in $A(X)^n$ at which there is a strong Condorcet winner and $\mathcal{L}_C = \wp_C \cap \mathcal{L}$ denote the set of all profiles of linear orders at which there is a strong Condorcet winner. The sets $\wp_C$ and $\mathcal{L}_C$ will both be referred to as the *Condorcet domain*, when doing so will not create ambiguity. Define the *Condorcet section of* $x \in X$ as the set $C_x \subset \mathcal{L}_C$ for which alternative $x \in X$ is the strong Condorcet winner.

2.2.2 Weak Non–Reversal

Eliaz (2004) provides a general impossibility theorem that encompasses those of Arrow (1951), Gibbard (1973), and Satterthwaite (1975) by defining a class of functions called *social aggregators*, which includes as special cases both social choice and welfare rules. The result hinges on the property of *preference reversal*, introduced by Eliaz (2004).

Definition 2.2.1 (Eliaz, 2004)**.** *A social welfare rule g with domain $\wp$ satisfies preference reversal if for every $x, y \in X$ and every two profiles $p, q \in \wp$, if $x \succ_{g(p)} y$ and $x \succ_i^p y$ implies $x \succ_i^q y$ for all $i \in N$, then $x \succ_{g(q)} y$.*

Campbell and Kelly (2006) utilize a similar definition in establishing a condition on social welfare functions that is necessary and sufficient for the social choice rule induced by selection the top–ranked alternative from the social ordering to be invulnerable to manipulation by coalitions, including singleton coalitions.

Definition 2.2.2 (Campbell and Kelly, 2006)**.** *A social welfare rule g with domain $\wp$ satisfies weak non–reversal if for every $x, y \in X$ and every profile $p \in \wp$, $x \succ_{g(p)} y$ implies that $x \succeq_{g(q)} y$ at all profiles $q \in \{s \in \wp \,:\, \forall i \in N,\ s(i) \neq p(i) \Rightarrow y \succ_i^s x\}$.*

Campbell and Kelly (2006) show that preference reversal is strictly more demanding than weak non–reversal, and note that weighted majority rule satisfies weak non–reversal but that the Borda rule does not when $\#X \geq 4$.

Within the social choice framework, the non–reversal condition of Campbell and Kelly (2006) reduces to the following definition.

Definition 2.2.3. *A social choice rule $g : \wp \to X$ satisfies **weak non–reversal** if for every $x, y \in X$ and every profile $p \in \wp$, $g(p) = x$ implies that $g(q) \neq y$ at all profiles $q \in \{s \in \wp : \forall i \in N,\ s(i) \neq p(i) \Rightarrow y \succ_i^s x\}$.*

From the statement of the definition, it is immediately clear that when the range of a social choice rule contains only two alternatives, weak non–reversal and strategy–proofness are logically equivalent. The notion of weak non–reversal is used in the characterization results of Section 2.4 to link the results of this paper to previous axiomatic characterizations of strategy–proofness.

2.3 Two Examples

The following example illustrates that the characterization of majority rule due to Campbell and Kelly (2003) does not hold when there is an even number of individuals.

Example 2.3.1. *Let there be an even number of individuals with strict preferences over the set of alternatives $X = \{x, y, z\}$. Define social choice rule g on the Condorcet domain $\mathcal{L}_C$ on X as follows: if alternative $x \in X$ is the Condorcet winner at profile $p \in \mathcal{L}_C$, then $g(p) = y$; if $y \in X$ is the Condorcet winner, then $g(p) = z$; if $z \in X$ is the Condorcet winner, then $g(p) = x$.*

It is clear that g is neither majority nor dictatorial rule and does not satisfy unanimity. Moreover, g cannot be manipulated: since n is even, an individual cannot unilaterally change the strong Condorcet winner within the domain $\mathcal{L}_C$, since a single individual cannot reduce a strict majority to a minority when there is an even number of individuals.

Note that the social choice rule in Example 2.3.1 is defined over the Condorcet domain for an odd number of individuals, but is not strategy–proof

The next example shows that requiring a social choice rule to satisfy unanimity is not sufficient to recover the characterization of majority rule when n is even.

Example 2.3.2. *Let there be an even number of individuals* ($n \geq 4$) *with strict preferences over the set of alternatives* $X = \{x, y, z\}$. *Define social choice rule* g *on the Condorcet domain* $\mathcal{L}_C$ *on* X *as follows: if alternative* $x \in X$ *is the Condorcet winner at profile* $p \in \mathcal{L}_C$, $g(p) = p_1(1)$*; if* $y \in X$ *is the Condorcet winner,* $g(p) = p_1(2)$*; if* $z \in X$ *is the Condorcet winner,* $g(p) = p_1(3)$.

Social choice rule g clearly satisfies unanimity and is non–dictatorial; it is also strategy–proof, appealing to the same logic as in Example 2.3.1.

The existence of strategy–proof social choice rules in Examples 2.3.1 and 2.3.2 hinges on the ability to define rules piece–wise on distinct Condorcet sections. Lemmas 2.4.2 and 2.4.3 establish this feature in greater generality, and are vital to the characterization of strategy–proof social choice rules in Theorem 2.4.4. It is worth noting that Examples 2.3.1 and 2.3.2 can be easily extended for any even $n \geq 4$ and are well–defined but not strategy–proof on the Condorcet domain with an odd number of individuals, as there exist preference profiles where individuals can misrepresent

their preferences to change the Condorcet winner.

2.4 Characterization Results

My main theorem provides a counterpart to the Campbell and Kelly (2003) result when n is even. As suggested by Examples 2.3.1 and 2.3.2, the Condorcet domain has a natural partition that can be used to define piecewise strategy–proof social choice rules when there is an even number of individuals.

Lemma 2.4.1. *The set of Condorcet sections, $\mathcal{C}_X$, forms a partition of $\wp_C$.*

The result follows immediately from the definition, but is included in Appendix A.2 for completeness. By construction, there exists a strong Condorcet winner $x \in X$ at every preference profile $p \in \wp_x$; hence, the members of $\mathcal{C}_X$ cover $\wp_C$. Moreover, since the strong Condorcet winner is unique when it exists, it must be that the members of $\mathcal{C}_X$ are disjoint.

It is worth noting that this partition exists independently of the parity of the set of individuals. However, when there is an even number of individuals, preference profiles in distinct Condorcet sections must differ in the preferences of at least two individuals. The next lemma formalizes this observation.

Lemma 2.4.2. *Suppose that n is even and $p \in \mathcal{C}_x$ for some $x \in X$. Then for all $i \in N$ and $q \in \wp_C$ such that $q(j) = p(j)$ for all $j \in N \setminus \{i\}$, $q \in \mathcal{C}_x$.*

The proof of Lemma 2.4.2 is given in Appendix A.3. The lemma reduces the analysis of strategy–proofness over the entire Condorcet domain to analysis over arbitrary Condorcet sections, as the following lemma establishes.

Lemma 2.4.3. *If n is even, then a social choice rule $g : \wp_C \to X$ is strategy–proof if and only if the restriction $g|_{C_x}$ is strategy–proof for each $x \in X$.*

The proof of Lemma 2.4.3 is given in Appendix A.4. With these lemmas, I provide a characterization result.

Theorem 2.4.4. *If $n > 4$ is even and $\#X \geq 3$, then the social choice rule $g : \mathcal{L}_C \to X$ is strategy–proof if and only if*

(i) $g|_{C_x}$ satisfies non–reversal for every subdomain C_x for which $\#g(C_x) = 2$, and;

(ii) $g|_{C_x}$ is dictatorial over the alternatives in $g(C_x)$ for every subdomain C_x for which $\#g(C_x) \geq 3$.

Propositions 2.4.5, 2.4.6, and 2.4.7 taken collectively provide a proof of Theorem 2.4.4, appealing to Lemma 2.4.3. These propositions characterize, respectively, strategy–proof social choice rules over Condorcet sections with ranges under g of size one, two, or three or more.

Proposition 2.4.5. *Consider the social choice rule $g : \mathcal{L}_C \to X$. If $n > 4$ is even and $\#g(C_x) = 1$ for some $x \in X$, then $g|_{C_x}$ is strategy–proof.*

The proof of Proposition 2.4.5 is trivial, but is provided in Appendix A.5 for completeness. Clearly any social choice rule is strategy–proof on a domain over which it has a singleton range; this holds independently of whether $x \in C_x$ or $x \notin C_x$. It is worth noting that majority rule is comprised of this case, with the range over each Condorcet section being a singleton set containing the strong Condorcet winner.

Proposition 2.4.6. *Consider the social choice rule $g : \mathcal{L}_C \to X$. If $n > 4$ is even and $\#g(C_x) = 2$ for some $x \in X$, then $g|_{C_x}$ is strategy–proof if and only if g satisfies weak non–reversal.*

The proof is available in Appendix A.6, and the result obtains whether $x \in C_x$ or $x \notin C_x$. As noted in Section 2.2.2, this result applies more generally: weak non–reversal is equivalent to strategy–proofness over any subdomain with a two–element range.

Proposition 2.4.7. *Consider the social choice rule $g : \mathcal{L}_C \to X$. A social choice rule g over $\mathcal{C}_x$ with $\#g(C_x) \geq 3$ is strategy–proof if and only if $g|_{C_x}$ is dictatorial with respect to $g(C_x)$.*

Appendix A.7 contains the proof of Proposition 2.4.7. Sufficiency is obtained immediately, since dictatorial rule is strategy–proof. The proof of necessity requires consideration of two cases: when $x \in C_x$ and when $x \notin C_x$.

When $x \in C_x$, the structure and content of the proof is similar to that of Campbell and Kelly (2006), but differs nontrivially at several crucial steps. The full details are provided for completeness. When $x \notin C_x$, the proof consists of two steps. In the first step, the Gibbard–Sattherthwaite theorem is invoked over a subdomain of C_x to show that if g is strategy–proof on C_x, then $g|_{C_x}$ is be dictatorial with respect to $g(C_x)$. In the second step, the strategy–proofness of g on C_x is shown to imply that the dictator over this subdomain must in fact be the dictator over the entire Condorcet section.

The following corollary obtains immediately from Theorem 2.4.4 by requiring g to satisfy unanimity.

Corollary 2.4.8. *If $n > 4$ is even and $\#X \geq 3$, then the unanimous social choice rule $g : \mathcal{L}_C \to X$ is strategy–proof if and only if*

(i) $g(C_x) = \{x\}$ *for every subdomain C_x for which $\#g(C_x) = 1$;*

(ii) $g|_{C_x}$ *satisfies non–reversal with $x \in g(C_x)$ and $g|_{C_x}(p) = x$ at all unanimous $p \in C_x$ for every subdomain C_x for which $\#g(C_x) = 2$, and;*

(iii) $g|_{C_x}$ *is dictatorial with respect to alternatives in $g(C_x)$ for every subdomain C_x for which $\#g(C_x) \geq 3$.*

The proof of Corollary 2.4.8 follows directly from Theorem 2.4.4. Unanimity implies that $x \in C_x$, so that $g|_{C_x}(p) = x$ whenever $\#g(C_x) = 1$. When $\#g(C_x) \geq 2$, the results are fundamentally unchanged save for those implied directly by the additional assumption of unanimity.

Corollary 2.4.8 demonstrates that requiring g to satisfy unanimity does not meaningfully alter the family of strategy–proof social choice rules over the Condorcet domain with an even number of individuals. In particular, it is not sufficient to extend the results of Campbell and Kelly (2006), as demonstrated by Example 2.3.2. However, an analogous theorem obtains and is expressed concisely with the introduction of two new definitions.

Definition 2.4.9. *Let Π be a partition of $\mathcal{L}_C$. Then $P \in \Pi$ is a* ***dictatorial section*** *of social choice rule $g : \mathcal{L}_C \to X$ on domain $\mathcal{L}_C$ if P is a Condorcet section and $g|_P$ is dictatorial.*

Definition 2.4.10. *Let P be a dictatorial section of social choice rule $g : \mathcal{L}_C \to X$. Then g is* ***quasi–majority rule*** *if for all profile $p \in P$, if $g(p)$ is not the Condorcet*

winner at p then (i) $\#g(P) = 2$, one element of which is the Condorcet winner at p, and (ii) $g|_P$ satisfies non–reversal.

With these definitions in place, Theorem 2.4.11 contains the main result.

Theorem 2.4.11. *If $n > 4$ is even and the unanimous social choice rule $g : \mathcal{L}_C \to X$ is strategy–proof and has no dictatorial sections, then g is quasi–majority rule.*

The proof of Theorem 2.4.11 follows directly from Theorem 2.4.4 and Definitions 2.4.9 and 2.4.10 . If g has no dictatorial sections, the range of g over any Condorcet section C_x must be a singleton or contain exactly two elements. Unanimity requires that when $\#g(C_x) = 1$, $g(C_x) = \{x\}$. The definition of quasi–majority rule incorporates the non–reversal requirement from Corollary 2.4.8 when $\#g(C_x) = 2$.

2.5 Conclusion

Combining the results of this paper with those of Campbell and Kelly (2003), yields a complete theory of non–dictatorial strategy–proof social choice rules on the Condorcet domain for an arbitrary finite number of individuals. Theorem 2.4.11 provides an analogue to the main result of Campbell and Kelly (2006) and illustrates the limitation of their characterization of majority rule. As made evident by Example 2.3.2, the class of non–dictatorial strategy–proof social choice rules on the Condorcet domain is much larger when there are an even number of individuals than when there are an odd number, in which case the class consists solely of majority rule.

The disparity between the results of this paper and the results of Campbell and

Kelly (2003) are due to the Condorcet section partition in the case of even number of individuals. It is this partition that allows for many strategy–proof rules to exist, and it is the barriers to manipulation between Condorcet sections that cause the method of proof used in the odd case to fail when applied to the even case. I believe that it may be possible to unify the results with respect to different parities of individuals by expanding the domain. A strong candidate for expansion would be the inclusion of profiles at which there exist only a weak Condorcet winner; that is, an alternative that is not strictly defeated by any other alternative in a pair–wise run-off. As suggested by Campbell and Kelly (2003), another possibility is the admission of profiles at which individuals are indifferent between distinct alternatives. It remains an interesting and open question as to whether a general strategy–proofness characterization of majority rule can be obtained over the appropriate domain of preference.

Chapter 3

Costly Self–Discovery:

Learning in Quasi–Hyperbolic Agents

3.1 Introduction

Behavioral economics has focused on two main types of agents with time–inconsistent preferences: naifs and sophisticates. Both types of agents are modeled as a sequence of autonomous temporal selves indexed by time, with per–period utility functions characterized by a (β, δ) pair, where $\beta \in (0, 1)$ represents the agent's discounting of all future periods relative to the current period, and $\delta \in (0, 1)$ represents the agent's discounting between consecutive periods. Naïve agents are unaware of the dynamic inconsistency of their preferences, and believe that future–period selves will make utility maximizing decisions that coincide with the decision that the current period self would make for them; that is, they believe that future–period will behave as if $\beta = 1$. Hence, utility–maximizing naïve agents do not choose to employ costly commitment

mechanisms that restrict the behavior of future–period selves. Sophisticated agents, on the other hand, recognize that their preferences are dynamically inconsistent and that without the implementation of commitment mechanisms, future–period selves will make decisions that are suboptimal with regard to their current–period utility function.

In infinite–horizon models of costly procrastination, the standard result is for naïve agents to procrastinate indefinitely and for sophisticated agents to employ a commitment mechanism in which the task is accomplished in the first period. DellaVigna and Malmendier (2006) offer data on health club memberships that do not appear to support these results. In particular, they observe that

> *On average, 2.29 full months elapse between the last [instance of health club] attendance and contract termination for monthly members. [...] This lag is at least four months for 20 percent of the users.*

Additionally, they note that "low–attendance consumers delaying canceling [their membership]...despite small transaction costs." That is, agents for whom the maintenance of a health club membership is costly relative to their benefits from exercise (i.e., those who do not frequently use the facilities) maintain their membership for a significant number of periods. However, even these agents eventually terminate their membership; in their data on monthly membership uses, over 75% of low–frequency members terminated their membership within twenty–four periods; for comparison, roughly one third terminated membership within 12 periods, and over 60% terminated their membership within 18 periods. As DellaVigna and Malmendier (2006) note, "observed consumer behavior is difficult to reconcile with standard preferences and beliefs."

Where DellaVigna and Malmendier (2006) propose a model of overconfidence in future self–control, this paper proposes a model of learning in which a learning–naïve agent slowly moves from naïve–type optimization behavior to sophisticated–type optimization behavior by experiencing stochastic per period shocks to her utility functions. The learning–naïve agent is initially uncertain about the value of β that enters into the utility functions of her past– and future–period selves, but holds prior beliefs μ_1 about the value, $\hat{\beta}$. By observing the stochastic shocks and action choices of past selves, the learning–naïve agent is able to update her beliefs about the value of $\hat{\beta}$, the quasi–hyperbolic discount factor of her past– and future–period selves. Given sufficiently many periods and sufficient variation in the stochastic shocks, the equilibrium behavior of the learning–naïve agent eventually converges to that of the sophisticated agent: when the parameters of the model are such that the sophisticated agent would employ a commitment mechanism, the learning–naïve agent will eventually do the same.

Section 3.2 proceeds with a discussion of the state of the literature on time–inconsistent preferences and learning in behavioral agents, with a particular focus on the model of Ali (2011). Section 3.3 presents a simple behavioral learning model, and provides analytic comparisons of the equilibrium behavior of the standard naïve and sophisticated agents in relation to the newly defined learning–naïve agent. A numerical example is provided and simulated in Mathematica, to concretely illustrate the differences in equilibrium behavior between the three types of behavioral agents. Finally, Section 3.4 concludes.

3.2 Learning in Behavioral Agents

The bulk of the literature on time–inconsistent preferences has focused on two distinct types of agents: the sophisticated agent, who fully recognizes her future self–control problems, and the naïve agent, who completely fails to recognize her future self–control problems. Building on the notion of time–inconsistent preferences originally developed by Phelps and Pollak (1968) and later formalized by Laibson (1994) and Barro (1999), O'Donoghue and Rabin (2001) propose a model in which a new type of agent, a partially–naïve agent with quasi–hyperbolic discount parameters (β, δ), believes with probability one that her future selves will instead optimize with respect to the quasi–hyperbolic discount factor is $\hat{\beta} \in (\beta, 1)$. From this modeling perspective, the naïve agent can be viewed as putting unit probability on future selves optimizing with respect to $\hat{\beta} = 1$, whereas the sophisticated agent puts unit probability on the true value of the quasi–hyperbolic discount factor, β.

In the setting of O'Donoghue and Rabin (2001), the partially–naïve agent faces a choice problem in a setting in which procrastination abounds. In particular, they show that an agent may be more likely to procrastinate in pursuit of important goals than in pursuit of unimportant ones, since the decision to pursue a task is based on its long–run benefit whereas the decision to complete the task is based on its immediate cost. By restricting attention to a newly defined class of strategies denoted *perception–perfect strategies*, O'Donoghue and Rabin (2001) are able to show that in any fixed choice environment, severe procrastination can occur only with a non–negligible degree of naiveté; that is, when $\hat{\beta} \gg \beta$. However, for any degree of naiveté, there exists a choice environment in which the agent procrastinates severely.

These results generalize in a choice framework those of O'Donoghue and Rabin (1999), who showed that even mild self–control problems, that is, those for which $\beta \to 1$, can cause severe procrastination in the completely naïve agent but not in the completely sophisticated agent.

Assessing the state of the behavioral economics literature on time–inconsistent preferences, Fudenberg (2006) remarks,

> *I think that behavioral economics would be well served by concerted attempts to provide learning–theoretic (or any other) foundations for its equilibrium concepts. At the least, this process might provide a better understanding of when the currently used concepts apply.*

Seeking to satisfy this aim through a model in which an agent's beliefs about the utility maximization undertaken by her future–period selves are not exogenously given, Ali (2011) proposes a model in which an agent who is only partially aware of her future temptations learns of them over time. In particular, this framework is in contrast to the modeling of O'Donoghue and Rabin's (2001) partially–naïve agent, for whom the belief that future–period selves optimize with respect to $\hat{\beta}$ is exogenously given and fixed throughout the infinite–horizon decision problem. As suggested by Ali (2011),

> *endogenizing beliefs in this way allows one to pose and answer the question of whether and when sophistication closely approximates the decision maker's self–awareness once he has had many opportunities to learn.*

Rather than a setting of procrastination, Ali (2011) models an environment in which the dynamically inconsistent agent faces temptation. In particular, the agent is modeled in the dual–self Doer/Planner framework of Fudenberg and Levine (2006), with each period in the infinite horizon decision problem broken into two sub–periods. In the first sub–period, the forward–looking rational Planner chooses a menu from

which the Doer will select an alternative; in the second sub–period, the temptation–prone Doer observes a realization of a stochastic shock to the agent's utility function and then chooses from the menu selected by the Doer. Within this context, there is a direct tension between desire for flexibility and the desire for commitment: the Planner is uncertain about the extent to which the Doer is able to resist temptation, but must select the menu prior to the realization of the agent's period–specific stochastic taste shock.

Learning in Ali (2011) therefore requires costly experimentation, as the Planner only learns about the Doer's level of self–control when she chooses a flexible menu that exposes the Doer to temptation. This model implies that in the limit, the Planner can only have incorrect beliefs about the Doer's self–control when she believes the Doer to be more subject to temptation than she in truth is; the Planner cannot perpetually overestimate the Doer's ability to resist temptation. However, if the Planner becomes sufficiently pessimistic about the Doer's self–control problem, she may select a singleton menu that fully commits the Doer to a single alternative. Since learning does not occur when the Doer does not face temptation, overly pessimistic beliefs concerning the Doer's ability to resist temptation may exist in perpetuity so long as the Planner pre–commits the Doer via menu selection. Ali (2011) presents a sufficient condition on the set of partial commitments, or menu choices, under which learning engenders sophistication.

In the next section, a model of learning is presented in the context of a single self quasi–hyperbolic agent facing an infinite horizon decision problem. Whereas Ali (2011) focuses on an environment in which temptation is the primarily embodiment

of dynamically inconsistent preferences, the model presented in Section 3.3 presents an environment in which dynamic inconsistency presents itself as procrastination. Combining model elements from O'Donoghue and Rabin (2001) and Ali (2011), the model introduces a new type of agent, the *learning–naïve agent* with quasi–hyperbolic discount parameters (β, δ) and initial beliefs μ_1 about the value $\hat{\beta}$ of the quasi–hyperbolic discount factor used by her future–period selves. Unlike the modeling environment of the naïve and sophisticated agents, but similar to that of the dual self agent of Ali (2011), the learning–naïve agent faces an infinite–period decision problem with period–specific stochastic shocks to her utility. In the spirit of Bénabou and Tirole (2004), the learning–naïve agent looks to her own past behavior to infer how she is likely to behave in the future. Whereas the standard behavioral types and the partially–naïve agent agent of O'Donoghue and Rabin (2001) repeatedly fail to reconcile their past behavior with their beliefs and former predictions of said behavior, the learning–naïve agent utilizes implemented actions resulting from the stochastic shocks to her utility function to update her beliefs about the behavior of her future–selves.

An important features of the model is the quasi–Bayesian nature of the learning–naïve agent's beliefs and updating process; in particular, the agent does not take into consideration the learning of future–period selves when updating her current–period beliefs and optimizing her planned sequence of actions. In particular, this is in contrast to the self–signaling models of Bodner and Prelec (1997, 2002) and Bénabou and Tirole (2004), as the current–period learning–naïve agent does not take into consideration the impact that her actions will have on the beliefs of future–period

selves regarding her tendency to procrastinate. Similar to the Doer/Planner agent of Ali (2011), the learning–naïve agent cannot be perpetually overly optimistic about the procrastination problem faced by her future–period selves; Theorem 3.3.1 contains this primary result.

The models of Ali (2011) and Section 3.3 present an environment in which agents with time–inconsistent preferences are able to learn about their susceptibility to temptation and procrastination, respectively. Both models are agnostic about the implications of this context–dependent learning on self–awareness more broadly. In particular, does learning about one's inclination to procrastinate completing an important task confer learning about one's ability to resist temptation? On an even finer grid, if an individual learns that she is susceptible to the temptations of chocolate despite her resolve to eat healthily, does she also learn about her susceptibility to the temptation of cigarettes despite her resolve to quit smoking? The models of Ali (2011) and Section 3.3 aim at providing learning–theoretic justification for the frequently used equilibrium assumption of sophistication; in addition to developing more satisfying and generalizable models of learning for time–inconsistent agents, it remains an interesting and open question as to how learning about β in one choice domain impacts beliefs about β in other domains.

3.3 A Simple Self–Learning Model

In this section, a simple model of self–learning by quasi–hyperbolic agents is proposed in a setting wherein agents are tempted to procrastinate. The model is meant to illustrate the primary features that behavioral self–learning model must possess;

Subsection 3.3.6 concludes the section with a discussion of the difficult modeling issues inherent to self–learning.

3.3.1 Notation & Definitions

Consider an infinite–horizon decision problem in which an agent enrolled in a health club membership must choose an action from the set

$$A = \left\{ \begin{array}{c} \text{Exercise Membership (E)}, \\ \text{Maintain Membership (M)}, \\ \text{Terminate Membership (T)} \end{array} \right\}. \tag{3.1}$$

If the membership is terminated in period t, the decision set of every subsequent period is the null set and the agent's continuation payoff is assumed to be zero. For notational convenience, let $d^-(t) = (d_1, d_2, \ldots, d_{t-1})$ denote the sequence of realized actions terminating in period $t-1$, and let $d^+(t) = (d_t, d_{t+2}, \ldots)$ denote the sequence of planned actions beginning in period t.

Agents are modeled as a sequence of autonomous temporal selves indexed by time $t \in \{1, 2, 3, \ldots\}$ with quasi–hyperbolic utility functions

$$U_t(d^-(t), d^+(t)) \equiv [\pi(d_{t-1}) + C(d_t)] + \beta \left(\sum_{i=1}^{\infty} \delta^i \left[\pi(d_{t-1+i}) + C(d_{t+i}) \right] \right), \tag{3.2}$$

where $\beta, \delta \in [0, 1]$, and $\pi(d_t)$ and $C(d_t)$ are the agent's instantaneous benefit and cost

functions, defined by

$$\pi(d_t) = \begin{cases} b, & \text{if } d_t = E \\ 0, & \text{otherwise} \end{cases} \tag{3.3}$$

and

$$C(d_t) = \begin{cases} c_M, & \text{if } d_t = M \\ c_T, & \text{if } d_t = T \\ c_E, & \text{if } d_t = E. \end{cases} \tag{3.4}$$

In each period t, the current–period self choses an action plan $d^+(t)$ subject to her history of realized decisions $d^-(t)$ that solves

$$\max_{d^+(t)} \left\{ [\pi(d_{t-1}) + C(d_t)] + \beta \left(\sum_{i=1}^{\infty} \delta^i \left[\pi(d_{t-1+i}) + C(d_{t+i}) \right] \right) \right\}. \tag{3.5}$$

Membership is assumed to be automatically renewing, so that the agent incurs a loss $c_M < 0$ in every period during which the membership is maintained but unused. During periods when the membership is exercised, the agent incurs an immediate loss $c_E < c_M$ attributable to the membership fee and the time and energy costs of exercising. It is assumed that terminating the membership incurs an immediate loss of c_T, where $c_E < c_T < c_M < 0$, which is attributable to the time cost of filing the appropriate paperwork.

Exercising the membership is assumed to be beneficial, in that exercising in period t delivers a benefit $b > 0$ in period $t + 1$. Of particular interest is the parameter space in which $c_E + \delta b > 0$ but $c_E + \beta\delta b < c_M$. When these assumptions hold the observed equilibrium behavior of naïve and sophisticated agents differ, as the

following subsections illustrate.

3.3.2 Behavioral Equilibria for the Naïve Agent

Consider first a naïve agent whose period–t utility maximization problem during a period of active membership is given by

$$\max_{d^+(t)} \left\{ [\pi(d_{t-1}) + C(d_t)] + \beta \left(\sum_{i=1}^{\infty} \delta^i \left[\pi(d_{t-1+i}) + C(d_{t+i}) \right] \right) \right\}, \tag{3.6}$$

but who believes that future period selves will solve

$$\max_{d^+(t)} \left\{ [\pi(d_{t-1}) + C(d_t)] + \left(\sum_{i=1}^{\infty} \delta^i \left[\pi(d_{t-1+i}) + C(d_{t+i}) \right] \right) \right\}. \tag{3.7}$$

That is, the naïve agent incorrectly believes that her future selves maximize a purely hyperbolic utility function, when in fact they optimize the same quasi–hyperbolic utility function as the current–period self. If the agent has terminated her membership in a previous period, the maximization is trivial and the agent receives a continuation payoff of zero in every subsequent period.

In period–t, the agent believes that her future selves solve

$$\max_{d^+(\tau)} \left\{ [\pi(d_{\tau-1}) + C(d_\tau)] + \left(\sum_{i=1}^{\infty} \delta^i \left[\pi(d_{\tau-1+i}) + C(d_{\tau+i}) \right] \right) \right\}, \tag{3.8}$$

which is maximized by the future action plan $d^+(\tau) = (E, E, E, \dots)$ regardless of past actions, since $c_T < c_M < 0 < c_E + \delta b$ by assumption. Anticipating that future selves will choose to exercise the membership in every period regardless of her current

period action, the naïve period–t agent believes her optimization problem to collapse to

$$\max \ \{c_E + \beta\delta b, c_M, c_T\}, \tag{3.9}$$

corresponding to the utility realized from exercising, maintaining, or terminating membership in the current period, respectively. Since $c_T < c_M$ and $c_E + \beta\delta b < c_M$, the naïve agent maximizes anticipated utility in every period t by choosing the future action plan $d^+(t) = (M, E, E, \dots)$.

In this parameter space, the observed equilibrium behavior of a naïve agent exemplifies procrastination: in every period, the current–period self optimizes autonomously as a β–δ discounter by maintaining membership, believing that future–period selves will optimize as pure hyperbolic discounters and exercise the membership. For every $t < \infty$, the sequence of realized past actions is therefore given by $d^-(t) = (M, M, \dots)$. Therefore, the present–discounted equilibrium payoff to the naïve agent is given by

$$c_M + \beta \sum_{i=1}^{\infty} \delta^i c_M = \left(\frac{1 - \delta + \beta\delta}{1 - \delta} \right) c_M. \tag{3.10}$$

3.3.3 Behavioral Equilibria for the Sophisticated Agent

The sophisticated agent recognizes that in every period, her current–period self will solve the utility maximization problem

$$\max_{d^+(t)} \ \left\{ [\pi(d_{t-1}) + C(d_t)] + \beta \left(\sum_{i=1}^{\infty} \delta^i \left[\pi(d_{t-1+i}) + C(d_{t+i}) \right] \right) \right\}, \tag{3.11}$$

and that by allowing each future self to optimize autonomously, her flow–utility in every period will be $c_M < 0$, as in each period the current–period self will selfishly choose to forgo exercising or terminating in favor of maintaining the membership.

The sophisticated agent will therefore seek a commitment mechanism to restrict the behavior of future selves whenever such a mechanism can be employed at sufficiently low cost; within the context of the proposed health club membership model, membership termination serves as such a commitment device. Two classes of equilibria of are particular interest: pure–strategy alternating termination equilibria, in which the agent terminates in the first period along the equilibrium path but delays termination for k periods if termination does not occur as planned, and probabilistic per–period termination equalibria, in which the agent terminates membership with probability p in every period, conditional on not having terminated in the past.

Alternating Termination Equilibria (ATE). Suppose that the sophisticated agent, recognizing her procrastination problem, employs a strategy in which she terminates her membership immediately along the equilibrium path; if she does not terminate her membership in the first period, she maintains membership for k periods, at which point the strategy repeats. Under such an equilibrium it is assumed that the informational structure is such that every agent knows the action prescribed to them by the equilibrium, so that if the equilibrium strategy prescribes to the first period self membership termination but to the second period self membership maintenance, the second period self knows that she is to maintain membership should she have the opportunity to play. For the proposed equilibrium behavior to be incentive compatible for the first–period self to whom the proposed strategy assigns the action

of terminating membership, the delay length k must be chosen so that

$$c_T \geq c_M + \beta \left[\sum_{i=1}^{k} \delta^i c_M + \delta^{k+1} c_T \right], \tag{3.12}$$

which implies that the payoff to terminating immediately is greater than the payoff to maintaining the membership for k additional periods before terminating. Moreover, the payoff to terminating immediately is greater than the payoff to exercising in the current period, given the equilibrium strategies of the future–period selves, since $c_E + \beta\delta b < c_M$ and Equation 3.12 imply that

$$c_T > c_E + \beta \left[\delta b + \sum_{i=1}^{k} \delta^i c_M + \delta^{k+1} c_T \right]. \tag{3.13}$$

Note that since the decision problem is stationary conditional on membership having never been terminated in the past, the above condition insures incentive compatibility in periods $t \in \{1, k+2, 2k+3, 3k+4, \dots\}$; that is, the above condition on k implies incentive compatibility for each agent to whom the proposed equilibrium strategy assigns the action of terminating membership.

Incentive compatibility of the proposed equilibrium in periods immediately following a period in which termination is prescribed by the equilibrium requires that k be such that

$$c_M + \beta \left[\sum_{i=1}^{k-1} \delta^i c_M + \delta^k c_T \right] \geq c_T, \tag{3.14}$$

so that the payoff to terminating immediately is less than the payoff to maintaining the membership for $k-1$ additional periods. Note that, as before, the assumption that $c_E + \beta\delta b < c_M$ insures that the payoff to exercising in the current period is also

less than the payoff to maintaining membership. Satisfaction of the incentive compatibility constraints in periods immediately following those in which termination is prescribed guarantees satisfaction of these constraints in all periods when membership maintenance is prescribed: since the number of periods until termination is declining after each period following a period of prescribed termination, the implied cost of further delaying termination is highest for the agent who acts in the period immediately following a period of prescribed termination. From Equation 3.14 this implication is obvious, since $c_T < c_M < 0$ and the length of the summation decreases for each subsequent agent in the maintenance sequence. Since the cost of terminating is fixed and the cost of delaying termination is therefore declining, any k that satisfies Equation 3.14 will satisfy the incentive compatibility constraints of all other agents to whom the proposed equilibrium strategy prescribes maintaining the membership.

Following the previous paragraphs, an alternating termination equilibrium exists for the sophisticated agent whenever there exists $k \in \mathbb{N}$ satisfying

(1) $c_T \geq c_M + \beta \sum_{i=1}^{k} \delta^i c_M + \beta\delta^{k+1} c_T$, and

(2) $c_T \leq c_M + \beta \sum_{i=1}^{k-1} \delta^i c_M + \beta\delta^{k} c_T$.

In specific applications, the existence of such equilibria could be verified numerically from assumed or estimated termination and maintenance costs, and the discount factors β and δ.

In an alternating termination equilibrium as described above, the sophisticated agent terminates her membership immediately along the equilibrium path. Therefore, the present–discounted payoff to the sophisticated agent resulting from this equilib-

rium strategy is given by c_T. The agent will therefore employ membership termination via an alternating termination equilibrium as a commitment mechanism whenever

$$c_T \geq \left(\frac{1 - \delta + \beta\delta}{1 - \delta} \right) c_M. \tag{3.15}$$

Note from Equation 3.26 that the sophisticated agent is more likely to utilize the commitment mechanism whenever (1) $c_T < 0$ increases, so that the immediate cost of terminating declines; (2) $c_M < 0$ decreases, so that the per–period cost of not terminating increases; (3) β or δ increase, so that the agent becomes more patient. See Appendix B.1 for verification of the relationship between commitment mechanism usage and patience.

Probabilistically Terminating Equilibria (PTE). Suppose that the sophisticated agent, recognizing her procrastination problem, instead employs a strategy in which she terminates her membership with some fixed probability $p \in (0, 1)$ in every period, conditional on never having terminated her membership in the past. For the proposed equilibrium to be incentive compatible, it must be that the expected payoff to maintaining membership in the current period is equal to the expected payoff to terminating in the current period. That is, p must be chosen so that

$$c_T = c_M + \beta \sum_{i=1}^{\infty} \delta^i (1 - p)^i [p c_T + (1 - p) c_M], \tag{3.16}$$

or, equivalently, to solve the quadratic

$$\beta\delta(c_T - c_M)p^2 + \delta(c_T - c_M - \beta[c_T - 2c_M])p + (1 - \delta)(c_T - c_M) - \beta\delta c_M = 0. \tag{3.17}$$

Since $c_E + \beta\delta b < c_M$ by assumption, it follows that the current–period agent cannot profitably deviate to any strategy that puts positive weight on exercising in the current period, given the equilibrium strategies of future selves. That is, the current–period self cannot profitably deviate from the proposed equilibrium strategy. As with the alternating termination equilibria, the existence of a probabilistic terminating equilibrium could be verified in a specific application by numerically calculating p in Equation 3.16 from assumed or estimated termination and maintenance costs, and the discount factors β and δ.

In a probabilistically terminating equilibrium as described above, the sophisticated agent terminates her membership after a finite number of periods along the equilibrium path, where the number of periods until termination is distributed geometrically with parameter p. Therefore, the present–discounted expected payoff to the sophisticated agent resulting from this equilibrium strategy is given by

$$pc_T + (1-p)c_M + \beta \sum_{i=1}^{\infty} \delta^i (1-p)^i [pc_T + (1-p)c_M]. \tag{3.18}$$

The agent will therefore employ membership termination via a probabilistically terminating equilibrium as a commitment mechanism whenever

$$pc_T + (1-p)c_M + \beta \sum_{i=1}^{\infty} \delta^i (1-p)^i [pc_T + (1-p)c_M] \geq \left(\frac{1-\delta+\beta\delta}{1-\delta} \right) c_M, \tag{3.19}$$

where p is chosen so as to satisfy Equation 3.16.

3.3.4 Behavioral Equilibria for the Learning–Naïve Agent

In addition to the standard behavioral agent types (naifs and sophisticates), consider a *learning–naïve agent* endowed with beliefs concerning the utility function her past and future selves maximize. In particular, suppose that in the initial period the learning–naïve agent knows that her current period self maximizes utility with respect to the quasi–hyperbolic parameters (β, δ), but has beliefs μ about the value $\hat{\beta}$ of the quasi–hyperbolic discount factor entering into the utility calculation of her past– and future–selves. Note that the initial beliefs of a learning–naïve agent can be made arbitrary close to those of a naïve agent in the following sense: for any $\epsilon \in (0, 1)$ there exists μ with full support on $(0, 1)$ such that $1 - \epsilon < \mathbb{E}_{\mu}[\hat{\beta}] < 1$.

Unlike the modeling framework in which naïve and sophisticated agents reside, learning–naïve agents reside in a modeling framework that permits learning about the value of the quasi–hyperbolic discount factor that enters the utility calculations of past and future selves through stochastic shocks to the benefit of exercising membership services. At the beginning of each period, the learning–naïve agent realizes a next–period benefit to exercise b of the random variable B with probability density function f on $[0, \infty]$, so that the instantaneous benefit function is given by

$$\pi(d_t, b_t) = \begin{cases} b_t, & \text{if } d_t = E \\ 0, & \text{otherwise.} \end{cases} \tag{3.20}$$

These shocks can be though to reflect exogenous factors that effect the benefit to exercising membership services but do not influence membership maintenance or ter-

minations costs. In the case of health club membership, these fluctuations might reflect daily changes in diet and physical activity that compliment or substitute for the use of health club services.

The decision problem facing the learning–naïve agent therefore proceeds as follows: at the beginning of period t of active membership, the agent's beliefs are updated in a quasi–Bayesian fashion conditional on the observed action and stochastic benefit to exercising from the previous period, (d_{t-1}, b_{t-1}). The period–t agent believes that her previous period self would have exercised her membership if and only if the discounted benefits exceed the cost; that is, if and only if

$$c_E + \hat{\beta}\delta b_t \geq 0 \qquad \Leftrightarrow \qquad \hat{\beta} \geq \frac{-c_E}{\delta b_t}. \tag{3.21}$$

Therefore, having observed $d_{t-1} = E$, the period–t agent updates her beliefs about $\hat{\beta}$ according to

$$\begin{aligned} \mu_t(x) &\equiv \mu_{t-1}(x|d_{t-1} = E, b_{t-1}) \\ &= \begin{cases} 0, & \text{if } 0 \leq x \leq \dfrac{-c_E}{\delta b_{t-1}} \\ \dfrac{\mu_{t-1}(x)}{\displaystyle\int_{-c_E/\delta b_{t-1}}^{1} \mu_{t-1}(y)dy}, & \text{if } \dfrac{-c_E}{\delta b_{t-1}} < x \leq 1. \end{cases} \end{aligned} \tag{3.22}$$

Similarly, after observing $d_{t-1} = M$ the agent updates her beliefs according to

$$\begin{aligned} \mu_t(x) &\equiv \mu_{t-1}(x|d_{t-1} = M, b_{t-1}) \\ &= \begin{cases} \dfrac{\mu_{t-1}(x)}{\displaystyle\int_0^{-c_E/\delta b_{t-1}} \mu_{t-1}(y)dy}, & \text{if } 0 \leq x \leq \dfrac{-c_E}{\delta b_{t-1}} \\ 0, & \text{if } \dfrac{-c_E}{\delta b_{t-1}} < x \leq 1. \end{cases} \end{aligned} \tag{3.23}$$

Having updated her beliefs, μ_t, the agent observes a realization b_t of the next–period benefits to exercising her membership and then maximizes her current period utility by solving

$$\max_{d^+(t)} \left\{ [\pi(d_{t-1}, b_{t-1}) + C(d_t)] + \beta \left(\sum_{i=1}^{\infty} \delta^i \left[\pi(d_{t-1+i}, \mathbb{E}_f[b_{t-1+i}]) + C(d_{t+i}) \right] \right) \right\} \tag{3.24}$$

subject to her belief that all future selves will maximize utility by solving

$$\max_{d^+(\tau)} \left\{ [\pi(d_{\tau-1}, b_{\tau-1}) + C(d_\tau)] + \mathbb{E}_{\mu_t}[\hat{\beta}] \left(\sum_{i=1}^{\infty} \delta^i \left[\pi(d_{\tau-1+i}, \mathbb{E}_f[b_{\tau-1+i}]) + C(d_{\tau+i}) \right] \right) \right\}, \tag{3.25}$$

where $\mathbb{E}_f[b_t]$ is the degenerate distribution that puts unit mass on the realized benefit b_t observed by the agent at the beginning of period t, and $\mathbb{E}_{\mu_\tau}[\hat{\beta}]$ is the expected value of the quasi–hyperbolic discount factor used by past– and future–period selves taken with respect to period–t believes μ_t.

When deciding whether or not to terminate membership in period–t, the learning–naïve agent compares her beliefs about the value of the quasi–hyperbolic discount factor used by past– and future–period selves to an exogenously determined benchmark

$\bar{\beta}$ derived from the observed equilibrium behavior of the sophisticated agent. When $\mathbb{E}_{\mu_t}[\hat{\beta}] < \bar{\beta}$, the learning–naïve agent terminates her membership, anticipating that future–selves will inefficiently maintain membership as a result of their (partially) unrecognized procrastination incentive. Natural benchmark β–values are those at which the sophisticated agent is indifferent between utilizing a commitment mechanism via either an alternating termination equilibrium or a probabilistically terminating equilibrium. In the health club membership model, these benchmarks are given by $\bar{\beta}_A$ and $\bar{\beta}_P$ respectively, and following Equations 3.26 and 3.19 are given by

$$c_T = \left(\frac{1 - \delta + \bar{\beta}_A \delta}{1 - \delta} \right) c_M. \tag{3.26}$$

and

$$pc_T + (1-p)c_M + \bar{\beta}_P \sum_{i=1}^{\infty} \delta^i (1-p)^i [pc_T + (1-p)c_M] = \left(\frac{1 - \delta + \bar{\beta}_P \delta}{1 - \delta} \right) c_M, \tag{3.27}$$

respectively.

It is important to note that the learning–naïve agent is not fully Bayesian in three fundamental respects. First, the agent fails to fully recognize that her past behavior is at odds with her beliefs about the utility functions of her past– and future–selves; this particular feature of the learning–naïve agent is also exhibited by the naïve agent who, despite a history of maintained membership, consistently maintains membership in accordance with her belief that her future selves will exercise membership. Secondly, the agent does not anticipate learning by her future selves when forecasting their utility–maximizing behavior. That is, she believes that all future–period

selves will believe that all other future–period selves will optimize with respect to quasi–hyperbolic discount factor $\mathbb{E}_{\mu_t}[\hat{\beta}]$, even though the agent will update her beliefs about the quasi–hyperbolic discount factor being used by future–period selves at the beginning of period–$(t+1)$. Finally, when choosing whether or not to terminate her membership, the learning–naïve agent ignores the option value of maintaining club membership for utilization during periods when the gains to exercising are exceptionally large. In particular, the agent uses benchmark β–values arising from the non–stochastic environment of the sophisticated agent to inform her decision in an environment with stochastic benefits.

Within this environment, the observed equilibrium behavior of the learning–naïve agent bridges those of the naïve and sophisticated agents when the parameters of the model are such that the sophisticated agent would utilize a commitment mechanism, as Theorem 3.3.1 provides.

Theorem 3.3.1. *For any μ_1 with full support on $(0,1)$, if f has full support on $\left(\frac{-c_E}{\beta\delta}, \frac{-c_E}{\delta}\right)$ and $\beta \leq \bar{\beta}$, the learning–naïve agent will terminate her membership in finite time.*

The proof of Theorem 3.3.1 follows in a straightforward fashion from well–known results about the convergence of beliefs under Bayesian updating, but available in Appendix B.2 for completeness. Theorem 3.3.1 provides conditions under which the equilibrium behavior of the learning–naïve agent will eventually mimic that of the sophisticated agent, for initial beliefs that are arbitrarily close to those of the naïve agent. That is, regardless of the initial level of naiveté, the learning–naïve agent learns of her tendency to procrastinate sufficiently quickly that she will optimally

terminate her membership within a finite number of periods with probability one. Subsection 3.3.5 presents a numerical example, in which the equilibrium behavior of the naïve and sophisticated agents can be directly compared to the stimulated behavior of the learning–naïve agent.

3.3.5 A Numerical Example

Following the empirical analysis of DellaVigna and Malmendier (2005), consider a health club in which members must pay a weekly membership fee of \$17 in order to have access to club equipment and services. Exercising costs an additional \$8 per week in time spent exercising and commuting, but delivers health benefits in the next week equivalent to \$60. The gym membership renews automatically, so that the agent is charged the monthly fee unless she choses to terminate her membership. Termination costs \$20 in time spent speaking to a customer service representative.[1]

Agents are assumed to have quasi–hyperbolic preferences, with parameters $\beta = \frac{1}{4}$ and $\delta = \frac{1}{2}$. In the notation of Subsection 3.3.1,

$$\pi(d_t) = \begin{cases} 60, & \text{if } d_t = E \\ 0, & \text{otherwise} \end{cases} \tag{3.28}$$

[1]In particular, DellaVigna and Malmendier (2005) report that the average monthly membership in their sample costs just under \$70 per month, or approximately \$17 per week. Additionally, they estimate a termination cost of at least \$15, which they attribute to the opportunity cost and mental cost of filing the appropriate paperwork. The costs in this example assume an additional transportation cost of \$3 per week to commute to and from the health club, as well as an additional \$5 per week opportunity cost for time spent exercising.

and

$$C(d_t) = \begin{cases} -17, & \text{if } d_t = M \\ -20, & \text{if } d_t = T \\ -25, & \text{if } d_t = E. \end{cases} \tag{3.29}$$

Moreover, the numerical values of the parameters meet the criteria established; that is,

$$c_E + \delta b \equiv -25 + \frac{1}{2}(60) = 5 > 0 \tag{3.30}$$

and

$$c_E + \beta\delta b \equiv -25 + \frac{1}{4}\left(\frac{1}{2}\right)(60) = -17.5 < -17 \equiv c_M. \tag{3.31}$$

Following the derivations of Subsection 3.3.2, in equilibrium the naïve agent maintains membership in every period, earning a present–discounted equilibrium payoff of

$$\left(\frac{1-\delta+\beta\delta}{1-\delta}\right) c_M = -\$21.25. \tag{3.32}$$

As in Subsection 3.3.3, the sophisticated agent may display two distinct equilibrium behaviors. An alternating termination equilibria with delay length k exists if an integer k can be chosen such that

$$\begin{aligned} c_T &\geq c_M + \beta \sum_{i=1}^{k} \delta^i c_M + \beta\delta^{k+1} c_T \\ \Leftrightarrow \quad -20 &\geq \frac{1}{4}(72^{-k} - 85) \\ \Leftrightarrow \quad -0.376 &\leq k, \end{aligned} \tag{3.33}$$

and

$$\begin{aligned} & c_T \leq c_M + \beta \sum_{i=1}^{k-1} \delta^i c_M + \beta \delta^k c_T \\ \Leftrightarrow \quad & -20 \leq 72^{-k-1} - \frac{85}{4} \\ \Leftrightarrow \quad & -1.052 \geq k. \end{aligned} \tag{3.34}$$

Since there exists no $k \in \mathbb{N}$ satisfying both of the inequalities above, an alternating termination equilibrium does not exist for the sophisticated agent for the parameters of this problem. A probabilistically terminating equilibrium exists if there is a termination probability $p \in (0, 1)$ that solves

$$\begin{aligned} c_T &= c_M + \beta \sum_{i=1}^{\infty} \delta^i (1-p)^i [p c_T + (1-p) c_M] \\ \Leftrightarrow \quad -20 &= -\frac{95 + 34p + 7p^2}{4 + 4p} \\ \Leftrightarrow \quad p &= \frac{23 - 2\sqrt{106}}{7} \approx 0.344. \end{aligned} \tag{3.35}$$

Therefore, within the numerical example given at the beginning of this section, the sophisticated agent displays a probabilistically terminating equilibria in which she terminates her membership with probability $p \approx 0.344$ in every period conditional on never having terminated in the past, earning a present–discounted expected equilibrium payoff of

$$p c_T + (1-p) c_M + \beta \sum_{i=1}^{\infty} \delta^i (1-p)^i [p c_T + (1-p) c_M] = \frac{52\sqrt{106} - 2943}{119} \approx -\$20.23. \tag{3.36}$$

Note that, as expected, the sophisticated agent is able to guarantee herself a payoff that exceeds that of the naïve agent by availing herself to the commitment mechanism

embodied by membership termination.

Following Subsection 3.3.4, consider a learning–naïve agent with true quasi–hyperbolic utility parameters $\beta = \frac{1}{4}$ and $\delta = \frac{1}{2}$, but with initial beliefs μ_1 about $\hat{\beta}$, the quasi–hyperbolic utility parameter that enters the utility function of her past and future selves. For simplicity, suppose that

$$\mu_1(x) = \begin{cases} \frac{1}{100}, & \text{if } 0 \leq x \leq \frac{98}{99} \\ \frac{4901}{50}, & \text{if } \frac{98}{99} < x \leq 1, \end{cases} \tag{3.37}$$

for $x \in [0, 1]$, which implies that $\mathbb{E}_{\mu_1}[\hat{\beta}] = 0.99$ so that the learning–naïve agent has initial beliefs very similar to those of a naïve agent. Moreover, suppose that stochastic shocks to the benefit of exercising membership services are distributed according to the shifted exponential distribution,

$$f(x) = \frac{1}{10} \exp\left[\frac{50 - x}{10}\right]. \tag{3.38}$$

Given the distribution of shocks, having observed $d_{t-1} = E$, the period–t learning–naïve agent updates her beliefs about $\hat{\beta}$ at the beginning of period–t according to

$$\begin{aligned} \mu_t(x) &\equiv \mu_{t-1}(x|d_{t-1} = E, b_{t-1}) \\ &= \begin{cases} 0, & \text{if } 0 \leq x \leq \frac{50}{b_{t-1}} \\ \dfrac{\mu_{t-1}(x)}{\displaystyle\int_{50/b_{t-1}}^{1} \mu_{t-1}(y)dy}, & \text{if } \frac{50}{b_{t-1}} < x \leq 1. \end{cases} \end{aligned} \tag{3.39}$$

Similarly, after observing $d_{t-1} = M$ the agent updates her beliefs according to

$$\begin{aligned} \mu_t(x) &\equiv \mu_{t-1}(x|d_{t-1} = M, b_{t-1}) \\ &= \begin{cases} \dfrac{\mu_{t-1}(x)}{\displaystyle\int_0^{50/b_{t-1}} \mu_{t-1}(y)dy}, & \text{if } 0 \leq x \leq \frac{50}{b_{t-1}} \\ \\ 0, & \text{if } \frac{50}{b_{t-1}} < x \leq 1. \end{cases} \end{aligned} \tag{3.40}$$

Having updated her beliefs, the learning–naïve agent observes a draw b_t from B and maximizes her expected utility by solving

$$\max_{d^+(t)} \left\{ [\pi(d_{t-1}, b_{t-1}) + C(d_t)] + \frac{1}{4} \left(\sum_{i=1}^{\infty} \frac{1}{2^i} \left[\pi(d_{t-1+i}, \mathbb{E}_f[b_{t-1+i}]) + C(d_{t+i}) \right] \right) \right\} \tag{3.41}$$

subject to her belief that all future selves will maximize utility by solving

$$\max_{d^+(\tau)} \left\{ [\pi(d_{\tau-1}, b_{\tau-1}) + C(d_\tau)] + \mathbb{E}_{\mu_t}[\hat{\beta}] \left(\sum_{i=1}^{\infty} \frac{1}{2^i} \left[\pi(d_{\tau-1+i}, \mathbb{E}_f[b_{\tau-1+i}]) + C(d_{\tau+i}) \right] \right) \right\}, \tag{3.42}$$

where $\mathbb{E}_f[b_t]$ is the degenerate distribution that puts unit mass on the realized benefit b_t observed by the agent at the beginning of period t, and $\mathbb{E}_{\mu_\tau}[\hat{\beta}]$ is the expected value of the quasi–hyperbolic discount factor used by past– and future–period selves taken with respect to period–t believes μ_t.

When deciding whether or not to terminate membership, the learning–naïve agent compares her beliefs about the value of the quasi–hyperbolic discount factor used by past– and future–period selves to an exogenously chosen benchmark derived from the observed equilibrium behavior of the sophisticated agent. Since an alternating

termination equilibrium does not exist for the parameters of this numerical example, the relevant benchmark is the β–value at which sophisticated agent is indifferent between employing a probabilistically terminating equilibrium to serve as a commitment mechanism and incurring the cost of indefinitely maintaining membership. That is, the relevant benchmark against which to compare $\mathbb{E}_{\mu_\tau}[\hat{\beta}]$ is given by

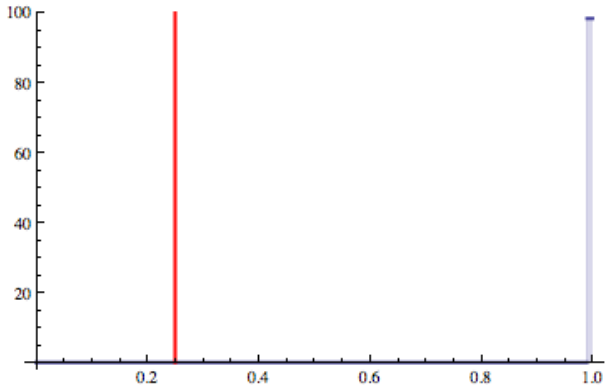
$$\begin{aligned} pc_T + (1-p)c_M + \bar{\beta}_P \sum_{i=1}^{\infty} \delta^i (1-p)^i [pc_T + (1-p)c_M] &= \left(\frac{1 - \delta + \bar{\beta}_P \delta}{1-\delta} \right) c_M \\ \Leftrightarrow \qquad \frac{3456 - 278\sqrt{106} + 4(59\sqrt{106} - 535)\bar{\beta}_P}{7(\sqrt{106} - 15)} &= -17 \left(\frac{1 + \bar{\beta}_P}{4} \right) \\ \Leftrightarrow \qquad \bar{\beta}_P &= \frac{3(53\sqrt{106} - 557)}{5(71\sqrt{106} - 785)} \\ &\approx 0.126. \end{aligned} \tag{3.43}$$

where p is given by Equation 3.35.

The stochastic environment of the learning–naïve agent was simulated in Mathematica, and the simulation code is available in Appendix B.3. Averaging over 1,000 simulation runs, the learning–naïve agent agent maintained her membership for an average of 180 but a median of only 3 periods before terminating. The learning–naïve agent therefore earns a present–discounted expected equilibrium payoff bounded below by

$$c_M + \beta \left[\delta c_M + \delta^2 c_M + \delta^3 c_T \right] = -\frac{333}{16} \approx -\$20.81, \tag{3.44}$$

which is less than the present–discounted expected equilibrium payoff to the sophisticated agent under the probabilistically terminating equilibrium but greater than the present–discounted equilibrium payoff to the naïve agent. Depending on the



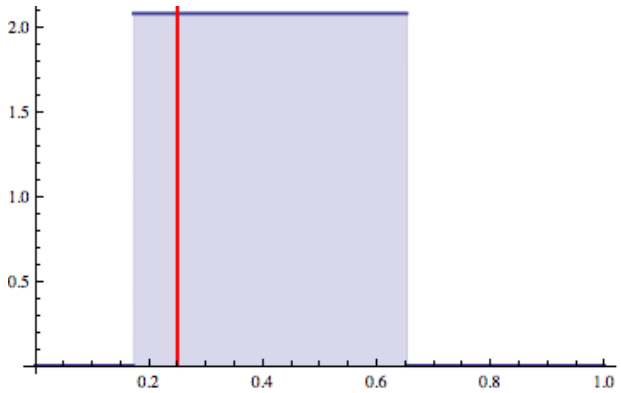
$$\mu_1(x) = \begin{cases} 0.01, & 0 \leq x \leq 0.99 \\ 98.02, & 0.99 < x \leq 1 \end{cases}$$

Figure 3.1: Probability Density Function for Simulated $\mu_1(x)$.

implemented per–period actions and the realized stochastic benefit to exercising, the realized present–discounted expected equilibrium payoff of the median learning–naïve agent is unbounded, since the distribution of benefits has infinite support. Tables 3.3.5–3.3.5 illustrate the path of beliefs for single simulation, in which eight periods passed before termination. The benchmark $\hat{\beta}$ is illustrated in red.

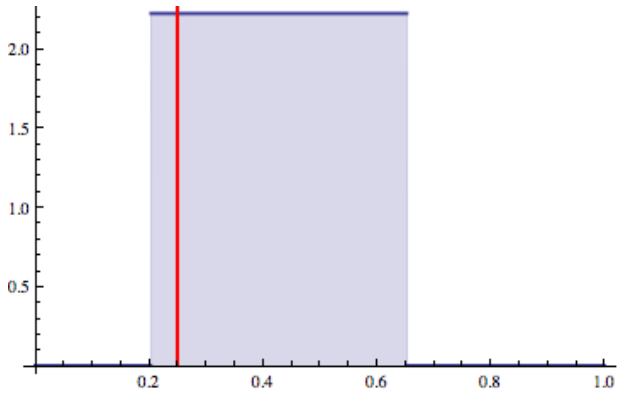
3.3.6 Modeling Issues & Alternative Assumptions

Like many of the behavioral learning models discussed in Section 3.2, the model of Section 3.3 is built upon several assumptions with which ideal models could dispense without unduly compromising the analytic tractability or obscuring the intuition of the model. First and foremost, the period–t learning–naïve agent forgets the parameters of the utility function maximized in the previous period: despite the fact that the true value of β entered the utility calculation of the period–$(t-1)$ self, the current–period self believes all past–selves to have maximized with respect to $\mathbb{E}_{\mu_t}[\hat{\beta}]$. This



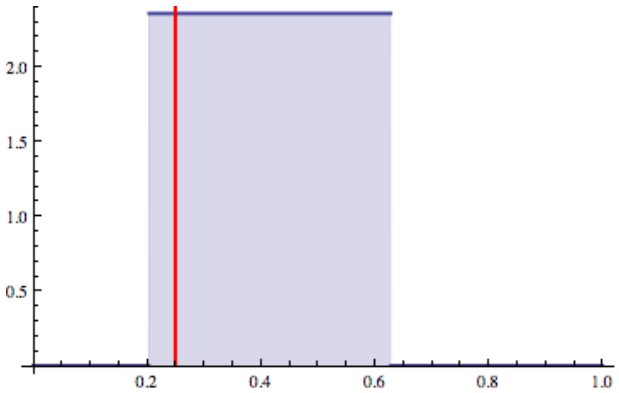
$$\mu_2(x) = \begin{cases} 0, & 0 \leq x \leq 0.17 \\ 2.08, & 0.17 < x \leq 0.65 \\ 0, & 0.65 < x \leq 1 \end{cases}$$

Figure 3.2: Probability Density Function for Simulated $\mu_2(x)$.



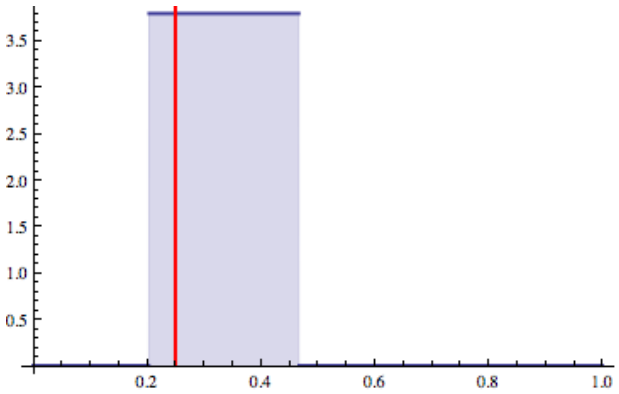
$$\mu_3(x) = \mu_4(x) = \begin{cases} 0, & 0 \leq x \leq 0.2 \\ 2.22, & 0.2 < x \leq 0.65 \\ 0, & 0.65 < x \leq 1 \end{cases}$$

Figure 3.3: Probability Density Function for Simulated $\mu_3(x)$, $\mu_4(x)$.



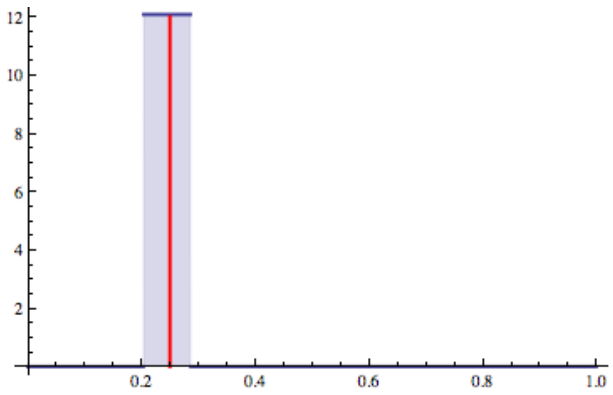
$$\mu_5(x) = \mu_6(x) = \begin{cases} 0, & 0 \leq x \leq 0.2 \\ 2.35, & 0.2 < x \leq 0.63 \\ 0, & 0.63 < x \leq 1 \end{cases}$$

Figure 3.4: Probability Density Function for Simulated $\mu_5(x)$, $\mu_6(x)$.



$$\mu_7(x) = \begin{cases} 0, & 0 \leq x \leq 0.2 \\ 3.79, & 0.2 < x \leq 0.47 \\ 0, & 0.47 < x \leq 1 \end{cases}$$

Figure 3.5: Probability Density Function for Simulated $\mu_7(x)$.



$$\mu_8(x) = \begin{cases} 0, & 0 \leq x \leq 0.2 \\ 12.07, & 0.2 < x \leq 0.29 \\ 0, & 0.29 < x \leq 1 \end{cases}$$

Figure 3.6: Probability Density Function for Simulated $\mu_8(x)$.

assumption is consistent with the fundamental assumption that the naïve agent, while maximizing in every period with respect to her true quasi–hyperbolic utility function, consistently ignores this fact in forecasting that future–selves with maximize a purely hyperbolic utility function.

When choosing whether or not to terminate her membership in period–t, the learning–naïve agent Section 3.3 ignores the option value of maintaining club membership for utilization in future periods in which the realized benefits to exercising membership are exceptionally large. Instead, the agent compares her estimate of the quasi–hyperbolic discount factor that enters into the utility function of future–period selves to an exogenously given benchmark value $\bar{\beta}$, and terminates her membership when $\mathbb{E}_{\mu_t}[\hat{\beta}] < \bar{\beta}$. The two benchmarks considered in Subsection 3.3.4 are given by those β–values at which the sophisticated agent would utilize a commitment mechanism as embodied in membership termination, under an alternating termination equilibrium and the probabilistically terminating equilibrium, respectively. In this sense,

the learning–naïve agent recognizes the imprecision in her beliefs and looks to the behavior of the sophisticated agent to inform her decision to terminate membership, despite the fact that the sophisticated agent operates in a deterministic environment completely lacking in the option value considerations inherent to the stochastic environment of the learning–naïve agent. From a modeling perspective, the motive to reduce the number of exogenously determined behavioral features is counterbalanced by the desire for tractability; incorporating considerations of option value into the learning–naïve agent's decision problem significantly complicates the model and introduces new philosophical tensions between the learning–naïve agent's apparent sophistication on some features of her decision problem and lack of sophistication on others.

As the learning–naïve agent is only able to learn by utilizing stochastic period–to–period variation in her decision problem, the source of this variation as presented in the model Subsection 3.3.4 provides a natural candidate for alternative modeling assumptions. In particular, rather than learning about her true tendency to procrastinate through stochastic shocks to the benefit of exercising, the learning–naïve agent could alternatively learn in an environment in which the benefits to exercise are fixed but the costs are stochastic, or an environment in which the cost of termination varies stochastically period–to–period while the cost and benefit of exercising membership remain fixed. Holding the timing of observation of the realization fixed, the models in which the stochastic element enters through the cost or the benefit of exercising are isomorphic. Since membership termination is assumed to be permanent, a model of behavioral learning similar to that presented in Subsection 3.3.4 in

which the stochastic element enters through the termination cost would only have meaningful implications if the agent can observe the realized cost prior to deciding to terminate; such a model is also isomorphic to the model presented herein.

As presented, the model isolates the behavioral learning dynamics from strategic experimentation by allowing the learning–naïve agent to view the realized next–period benefits to exercising before she commits to a particular action in the current period. By adjusting the timing of this realization, the structure of the model under each of the stochastic environments described above would differ substantially. In particular, if the realization of the stochastic benefits to exercising are observed by the agent only after she commits to exercising, the agent's decision problem becomes more stationary in the sense that only the expectation $\mathbb{E}_f[b_t]$ is used to determine the optimal action in period–t. However, it also becomes a model in which strategic experimentation plays a role, as learning only occurs following periods in which the agent chose to exercise; whenever $d_{t-1} = M$, such a model would imply that $\mu_t = \mu_{t-1}$.

Fine–tuning of the nature of the stochastic element in the model of Subsection 3.3.4 might allow for more realistic or natural model environments, and should likely depend on the details of the environment to which the model is being applied. These modeling choices aside, the departures of the model from those of fully rational Bayesian agents remain. At the heart of the issue remains the tension between the realism of a model and its tractability, a conflict fundamental to economic modeling.

3.4 Conclusion

Building on the work of O'Donoghue and Rabin (2001) and Ali (2011), this paper presents a model of a learning–naïve agent facing an infinite–horizon choice problem in an stochastic setting, whose equilibrium behavior bridges those of the standard naïve and sophisticated agents. The self–learning model presented here is motivated in part by the call of Fudenberg (2006) for more learning–theoretic based justifications for the assumption of sophistication and in part by the discrepancies between the reported empirical behavior in DellaVigna and Malmendier (2006) and those predicted by the standard time–inconsistent types. The results of this model, as well as the philosophical underpinnings of its framework, compliment the pioneering work of Ali (2011) to adapt the modeling of time–inconsistent agents to a rigorous learning–theoretic approach. The temptation–prone Planner/Doer agent of Ali (2011) represents a counterpart and alternative modeling approach to the procrastination–prone quasi–hyperbolic discounter of the model presented herein.

In each period, the learning–naïve agent implements an action that is optimal with regard to her beliefs about the value of the quasi–hyperbolic discount factor that enters into the utility calculations of her future selves and to the realized next–period benefit to exercising her membership. To achieve a tractable model and isolate the process of learning from the well–understood dynamics of strategic experimentation and option value, the agent is assumed to be quasi–Bayesian; when maximizing her expected utility she fails to take into consideration the additional learning to which future–period selves will be privilege, as well as the option value of maintaining membership for exercising in periods with exceptionally high next–period

benefits. Through advanced modeling techniques, future models may be able to incorporate these features of the decision problem without considerably compromising the tractability of the learning problem.

As in Ali (2011), the learning–naïve agent becomes fully aware of her tendency to procrastinate for certain regions of the model's parameter space; namely, whenever the benchmark value $\beta \leq \bar{\beta}$, the learning–naïve agent becomes sufficiently convinced of her procrastination problem and terminates her membership in finite time. When such learning does not occur, the direction of the bias is towards pessimism: if $\bar{\beta} < \beta$, the learning–naïve agent will sub–optimally terminator her membership with positive probability. The combined conclusion of these two results is that in this framework, the learning–naïve agent cannot remain indefinitely optimistic about her procrastination problem.

The issue of self–learning in time–inconsistent agents remains an area ripe for theoretical development. In many modeling contexts, the assumption of sophistication is practically and conceptually appealing; the further application of learning–theoretic models to such agents may provide a stronger foundation on which such assumptions rest. Moreover, the issue of learning remains of supreme interest in its own right. Continued theoretical developments on this front have great potential for welfare enhancing policy recommendations as well as for clear and concise theoretical predictions.

Chapter 4

Networked One–to–One Matching

4.1 Introduction

In the standard one–to–one matching model of Gale and Shapley (1962), agents are implicitly assumed to have complete knowledge of the identities of the other agents in the market, their own preferences over these agents, and the preferences of the other agents.

In this paper, the standard one–to–one two–sided matching model is augmented by introducing a concept of local stability. Notions of distance are encoded in a social network, in which directly linked agents are viewed as being "close" to one another. This network structure allows for the introduction of two new definitions of stable matchings, depending on the interpretation of the network. Under the first definition, the network is seen as limiting the set of potential blocking pairs: only agents who are linked in the network share enough information about one another's preferences to block a potential match. Under the second definition, the network is

seen as additionally imposing direct restrictions on the pairs of agents who can be matched: agents who are not located sufficiently close enough – that is, who are not linked – cannot be matched to one another at any stable matching.

Following Roth and Sotomayor (1990), the analogous properties of these two sets of network–stable matchings are derived for comparison to the standard model. Through the implementation of an augmented Deferred Acceptance Algorithm (Gale and Shapley, 1962), the existence of both types of network–stable matching is shown for generic marriage networks. Moreover, nesting relations between the two new network–stable matching sets and the set of stable matchings on the associated marriage problem are derived, and examples are provided to illustrate cases in which nesting does not generally hold. Finally, the marriage network framework is imagined to be preceded by a network formation game, in which agents simultaneously propose links while only having access to limited information about the market.

The model presented here differs in many crucial respects from the literature on strategic network formation. Jackson and Wolinsky (1996) model a dynamic setting in which the network structure evolves as agents reoptimize their sets of connections to maximize utility arising from communication across the network, and show that efficient networks need not be stable in such an environment. Using an implementation approach, Dutta and Mutuswami (1997) show that this tension can be partially reconciled in certain settings. Bala and Goyal (1999) characterize the architecture of equilibrium networks in a setting in which agents receive direct benefits from connecting to other agents in a network, and receive indirect benefits from neighbors' neighbors. Analyzing a similar model, Jackson and Rogers (2005) show that the

equilibrium network structures exhibit a certain "small worlds" property, in which densely connected groups of individuals are connected to other groups by sparse links. Bridging the gap between the literatures on network formation games and bargaining on networks, Bloch and Jackson (2007) consider a model with transferable utility in which agents propose and maintain links to maximize direct and indirect benefits received from their network connections. In contrast to these bodies of work, the model of marriage networks presented herein presents the network formation game as preceding a formal matching process, and as such provides microeconomic foundations for utility derived solely from network connections. Within the marriage networks framework, indirect benefits to connections are received insofar as they impact the set of network–stable matchings.

The marriage networks model presented here shares many common properties with that of Arcaute and Vassilvitskii (2009). Whereas this paper focuses on the theoretic properties of the sets of network–stable matchings, their work provides a reinterpretation of the Deferred Acceptance Algorithm of Gale and Shapley (1962) as a myopic best response dynamic. While providing a similar nesting result on one set of network–stable matchings and stable matchings to the associated marriage problem, Arcaute and Vassilvitskii (2009) approach the question from a computational perspective and show that proving the equality of these two sets is NP–complete.

In Section 4.2, the marriage problem is augmented to include a network structure and the new stability concepts of interested are introduced. The properties of these sets are derived and analyzed in Section 4.3, which includes several examples illustrating properties of stable matchings that do not translate to network–stable match-

ings. Following Gale and Shapley (1962), the existence of the relevant network–stable matchings are established for generic marriage networks, and the opposing interests of the two sides of the market are established, as in Knuth (1976). In Section 4.4, the marriage network framework is embedded in a network formation game. The natural definition of a Nash equilibrium in this network formation and matching game is introduced, and the set of such Nash equililbria is characterized in the full information setting. The section concludes with a discussion of the modeling issues inherent to network formation games and the likely intractability of a full characterization of the set of Nash equilibria in a private information setting.

4.2 Notation & Definitions

Consider two disjoint finite sets of agents, given by $\mathcal{M} = \{m_1, \ldots, m_M\}$ and $\mathcal{W} = \{w_1, \ldots, w_W\}$, which we will refer to as the sets of *men* and *women*, respectively. Let $\mathcal{N} = \mathcal{M} \cup \mathcal{W}$ be the set of all agents in the market. Each man $m \in \mathcal{M}$ is endowed with a strict preference relation $\succ_m$, which is a complete asymmetric binary relation on $\mathcal{W} \cup \{m\}$. Similarly, each woman $w \in \mathcal{W}$ is endowed with a strict preference relation $\succ_w$ over $\mathcal{M} \cup \{w\}$. Woman w is *unacceptable* to man m if $m \succ_m w$, and man m is unacceptable to woman w if $w \succ_w m$.

Let $\mathbf{P}_m$ denote the set of all preference relations for man m and define

$$\mathbf{P}_{\mathcal{M}} = \times_{m \in \mathcal{M}} \mathbf{P}_m, \tag{4.1}$$

with representative element $\succ_{\mathcal{M}} = (\succ_{m_1}, \ldots, \succ_{m_M})$. Preference domains for each

woman and the preference domain for the set of women are defined analogously. Define $\mathbf{P} = \mathbf{P}_{\mathcal{M}} \times \mathbf{P}_{\mathcal{W}}$, with representative element, or *preference profiles*,

$$\succ_{\mathcal{N}} = (\succ_{m_1}, \ldots, \succ_{m_M}, \succ_{w_1}, \ldots, \succ_{w_W}). \tag{4.2}$$

A *marriage problem* is a triple $\langle \mathcal{M}, \mathcal{W}, \succ \rangle$. Given a marriage problem, a *matching* μ is a 1–1 function $\mu : \mathcal{M} \cup \mathcal{W} \to \mathcal{M} \cup \mathcal{W}$ satisfying

(i) $\forall m \in \mathcal{M}, \forall w \in \mathcal{W}$, $\mu(m) = w$ if and only if $\mu(w) = m$;

(ii) $\forall m \in \mathcal{M}$, $\mu(m) \in \mathcal{W} \cup \{m\}$, and;

(iii) $\forall w \in \mathcal{W}$, $\mu(w) \in \mathcal{M} \cup \{w\}$.

The first requirement ensures that man m is matched with woman w if and only if woman w is matched with man m. The second and third requirements ensure that any man not matched to a woman is matched to himself (that is, he remains *unmatched*), and analogously for each woman.

For each agent $i \in \mathcal{N}$, let $\succsim_i$ be the weak preference relation on the set of matchings induced by $\succ_i$. That is, for matchings μ and ν, $\mu \succsim_i \nu$ if and only if $\mu(i) = \nu(i)$ or $\mu(i) \succ_i \nu(i)$.

Given a marriage problem $\langle \mathcal{M}, \mathcal{W}, \succ \rangle$, a matching μ is *individually rational* if for all $i \in \mathcal{N}$, $\mu(i) \succsim_i i$; that is, if no agent is matched to an agent he or she finds unacceptable. For a fixed matching μ, agents $(m, w) \in \mathcal{M} \times \mathcal{W}$ form a *blocking–pair of* μ if $w \succ_m \mu(m)$ and $m \succ_w \mu(w)$; that is, if agents m and w prefer to be matched to each other than to their assigned partners under μ. A matching μ is

stable for $\langle \mathcal{M}, \mathcal{W}, \succ \rangle$ if it is individually rational and there exist no blocking–pairs. Let $\mathbf{U}(\mathcal{M}, \mathcal{W}, \succ)$ denote the set of stable matchings on $\langle \mathcal{M}, \mathcal{W}, \succ \rangle$.

Having introduced the standard notations and definitions from matching theory, I next introduce the machinery needed to analyze one–to–one matching with an underlying network structure.

Definition 4.2.1. *A* ***marriage network*** *is a 4–tuple* $\langle \mathcal{M}, \mathcal{W}, \succ, \Gamma \rangle$*, where* $\Gamma \subseteq \mathcal{M} \times \mathcal{W}$.

Borrowing from the networks literature, the network Γ is a *null network* if $\Gamma = \emptyset$ and is *biclique* if $\Gamma = \mathcal{M} \times \mathcal{W}$. The network Γ can be viewed as encoding a binary relation on $\mathcal{N}$ that will be useful in defining stable matches on a marriage network. For the purposes of this section, we assume that the network structure Γ is exogenously given. In particular, it is not the result of a strategic network–formation process.

Definition 4.2.2. *Man* $m \in \mathcal{M}$ *and woman* $w \in \mathcal{W}$ *are* ***aquatinted*** *in* $\langle \mathcal{M}, \mathcal{W}, \succ, \Gamma \rangle$ *if* $(m, w) \in \Gamma$.

The acquaintance relation may be modeled as interacting with the set of stable matchings in two distinct ways: it could shrink the set of stable matchings by restricting which man–woman pairs are permitted under a matching, or it could expand the set of stable matchings by restricting which man–woman pairs are permitted to form blocking pairs. The next two definitions formalize these interpretations.

Definition 4.2.3. *Given a marriage network* $\langle \mathcal{M}, \mathcal{W}, \succ, \Gamma \rangle$*, a matching* μ ***respects*** Γ *if* $\mu(m) = w$ *implies that* $(m, w) \in \Gamma$.

Definition 4.2.4. *Given a marriage network* $\langle \mathcal{M}, \mathcal{W}, \succ, \Gamma \rangle$ *and a matching* μ*, agents* $(m, w) \in \mathcal{M} \times \mathcal{W}$ *form a* ***local blocking–pair in*** Γ *if* $w \succ_m \mu(m)$*,* $m \succ_w \mu(w)$*, and* $(m, w) \in \Gamma$*.*

With these definitions in mind, two new definitions of stable matchings on marriage networks can be introduced.

Definition 4.2.5. *Given a marriage network* $\langle \mathcal{M}, \mathcal{W}, \succ, \Gamma \rangle$*, the matching* μ *is* ***informationally*** Γ***–stable*** *if* μ *is individually rational and has no local blocking–pairs.*

Given a marriage network $\langle \mathcal{M}, \mathcal{W}, \succ, \Gamma \rangle$, let $\mathbf{U}_I^{\Gamma}(\mathcal{M}, \mathcal{W}, \succ)$ denote the set of informationally Γ–stable matchings. Informationally network–stable matchings are important in markets in which a centralized matching occurs, as the definition places no restrictions beyond individual rationality on which pairs of agents can be feasibly matched, but rather restricts the number of potential blocking pairs. The National Resident Matching Program (NRMP) represents a market in which informationally network–stable matchings are a relevant equilibrium concept: if residents and residency programs are unaware of the preferences of some of the other players in the market, any matching issued by the NRMP would be stable so long as it were individually rational and there were no resident–hospital pairs who were informed of each other's preferences and formed a local blocking pair. Hence, the market designer is constrained in choosing a matching by the local information available to the participants, but not by the existence of the network otherwise.

The second definition of stable matchings on marriage networks requires not only that a matching be individually rational and have no local blocking pairs, but also that it respect the network structure.

Definition 4.2.6. *Given a marriage network $\langle \mathcal{M}, \mathcal{W}, \succ, \Gamma \rangle$, the matching μ is **technologically** Γ**–stable** if μ respects Γ, is individually rational, and has no local blocking–pairs.*

Let $\mathbf{U}_T^{\Gamma}(\mathcal{M}, \mathcal{W}, \succ)$ denote the set of technologically Γ–stable matchings for a marriage network $\langle \mathcal{M}, \mathcal{W}, \succ, \Gamma \rangle$. In contrast to informationally network–stable matchings, technologically network–stable matchings are important in large markets in which matchings form without the aid of centralization. In such markets, agents can only form potential matchings with agents with whom they are acquainted. Consider, for example, the market for romantic partners in New York City. It is obvious that players are only informed of the existence of a small subset of the other players in the market, and of the preferences of an even smaller subset. Matchings can only occur between agents who know each other, and who prefer each other to any other agent with whom they are acquainted. Allowing the network structure to encode the set of potential partners with whom an agent is capable of matching, the network can be understood as representing a technological constraint. Examples of such technological constraints may include geographical distance, language barriers, or legislative restrictions on marriageable partners.

Given a marriage problem $\langle \mathcal{M}, \mathcal{W}, \succ \rangle$, agent j is *achievable* to agent i if there exists a stable matching μ under which $\mu(i) = j$. Similar definitions exist for the network stable matching concepts introduced above, and will be useful in analyzing the structure of the sets $\mathbf{U}_I^{\Gamma}(\mathcal{M}, \mathcal{W}, \succ)$ and $\mathbf{U}_T^{\Gamma}(\mathcal{M}, \mathcal{W}, \succ)$.

Definition 4.2.7. *Given a marriage network $\langle \mathcal{M}, \mathcal{W}, \succ, \Gamma \rangle$, agent j is **informationally achievable** to agent i if $\exists \mu \in \boldsymbol{U}_I^{\Gamma}(\mathcal{M}, \mathcal{W}, \succ)$ such that $\mu(i) = j$.*

Definition 4.2.8. *Given a marriage network $\langle \mathcal{M}, \mathcal{W}, \succ, \Gamma \rangle$, agent j is* ***technologically achievable*** *to agent i if $\exists \mu \in \boldsymbol{U}_T^{\Gamma}(\mathcal{M}, \mathcal{W}, \succ)$ such that $\mu(i) = j$.*

For notational convenience, let $\mathbf{A}(i; \mathcal{M}, \mathcal{W}, \succ)$ denote the set of agents who are achievable to agent i in the marriage problem $\langle \mathcal{M}, \mathcal{W}, \succ \rangle$, and let $\mathbf{A}_I^{\Gamma}(i; \mathcal{M}, \mathcal{W}, \succ)$ and $\mathbf{A}_T^{\Gamma}(i; \mathcal{M}, \mathcal{W}, \succ)$ denote the sets of agents who are, respectively, informationally and technologically achievable to agent i in the marriage network $\langle \mathcal{M}, \mathcal{W}, \succ, \Gamma \rangle$. These sets produce a natural relation on the set of agents $\mathcal{N}$, as the next lemma states.

Lemma 4.2.9. *Achievability induces a symmetric relation on $\mathcal{M} \times \mathcal{W}$; that is, $j \in \boldsymbol{A}_k^{\Gamma}(i; \mathcal{M}, \mathcal{W}, \succ)$ if and only if $i \in \boldsymbol{A}_k^{\Gamma}(j; \mathcal{M}, \mathcal{W}, \succ)$, for $k \in \{I, T\}$.*

The result follows directly from the definition of a network–stable matching: if there exists $\mu \in \mathbf{U}_k^{\Gamma}(\mathcal{M}, \mathcal{W}, \succ)$ such that $\mu(i) = j$, then necessarily $\mu(j) = i$, for $k \in \{I, T\}$. In Section 4.3, Corollary 4.3.2 will show that these sets have a natural nesting for any marriage network.

4.3 Properties of Network Stable Matchings

In this section, the properties and structure of informationally and technologically stable network matchings are considered. Whenever possible, results on marriage networks are compared and contrasted to well–established results on the analogous marriage problems.

The following lemma establishes a natural nesting of the sets of informationally and technologically network–stable matchings.

Lemma 4.3.1. *For any marriage network* $\langle \mathcal{M}, \mathcal{W}, \succ, \Gamma \rangle$*, the set of technologically* Γ*–stable matchings is contained within the set of informationally* Γ*–stable matching; that is,* $\boldsymbol{U}_T^\Gamma(\mathcal{M}, \mathcal{W}, \succ) \subseteq \boldsymbol{U}_I^\Gamma(\mathcal{M}, \mathcal{W}, \succ)$.

The proof follows from the definitions of informationally and technologically network–stable matchings. If $\mu \in \mathbf{U}_T^\Gamma(\mathcal{M}, \mathcal{W}, \succ)$, then by definition μ is individually rational and has no local blocking pairs. Hence, $\mu \in \mathbf{U}_I^\Gamma(\mathcal{M}, \mathcal{W}, \succ)$. Intuitively, technologically network–stable matchings are subject to the same blocking–pair constraints as informationally network–stable matchings, and are further subject to additional constraints concerning the set of feasible (that is, Γ respecting) matches.

Corollary 4.3.2. *For any marriage network* $\langle \mathcal{M}, \mathcal{W}, \succ, \Gamma \rangle$ *and any agent* $i \in \mathcal{N}$*, if agent* j *is technologically achievable to agent* i*, she is also informationally achievable to agent* i*; that is,* $\boldsymbol{A}_T^\Gamma(i; \mathcal{M}, \mathcal{W}, \succ) \subseteq \boldsymbol{A}_I^\Gamma(i; \mathcal{M}, \mathcal{W}, \succ)$.

The result follows from Lemmas 4.3.1: if there exists

$$\mu \in \mathbf{U}_T^\Gamma(\mathcal{M}, \mathcal{W}, \succ) \subseteq \mathbf{U}_I^\Gamma(\mathcal{M}, \mathcal{W}, \succ) \tag{4.3}$$

such that $\mu(i) = j$, then $j \in \mathbf{A}_T^\Gamma(i; \mathcal{M}, \mathcal{W}, \succ)$ and $j \in \mathbf{A}_I^\Gamma(i; \mathcal{M}, \mathcal{W}, \succ)$.

The next lemma establishes a natural nesting of the set of stable matchings on the marriage problem $\langle \mathcal{M}, \mathcal{W}, \succ \rangle$ and the set of informationally network–stable matchings on the associated marriage network $\langle \mathcal{M}, \mathcal{W}, \succ, \Gamma \rangle$.

Lemma 4.3.3. *For any marriage network* $\langle \mathcal{M}, \mathcal{W}, \succ, \Gamma \rangle$*, the set of stable matchings on* $\langle \mathcal{M}, \mathcal{W}, \succ \rangle$ *is contained within the set of informationally* Γ*–stable matchings on* $\langle \mathcal{M}, \mathcal{W}, \succ, \Gamma \rangle$*; that is,* $\boldsymbol{U}(\mathcal{M}, \mathcal{W}, \succ) \subseteq \boldsymbol{U}_I^\Gamma(\mathcal{M}, \mathcal{W}, \succ)$.

The proof follows from the definitions of stable matchings and informationally network–stable matchings. If $\mu \in \mathbf{U}(\mathcal{M}, \mathcal{W}, \succ)$, then by definition μ is individually rational and has no blocking pairs; in particular, this implies that μ has no local blocking pairs. Hence, $\mu \in \mathbf{U}_I^{\Gamma}(\mathcal{M}, \mathcal{W}, \succ)$. Intuitively, informationally network–stable matchings face fewer constraints in terms of potential blocking pairs than stable matchings.

Lemmas 4.3.1 and 4.3.3 provide weak nesting conditions; Lemma 4.3.4 gives conditions under which the sets of matchings under consideration coincide.

Lemma 4.3.4. *For any set of men $\mathcal{M}$ and any set of women $\mathcal{W}$, if network $\Gamma \subseteq \mathcal{M} \times \mathcal{W}$ is biclique, then for all preference profiles $\succ \in \boldsymbol{P}$, the sets of stable matchings, informationally network–stable matchings, and technologically network–stable matching coincide; that is, $\boldsymbol{U}(\mathcal{M}, \mathcal{W}, \succ) = \boldsymbol{U}_T^{\Gamma}(\mathcal{M}, \mathcal{W}, \succ) = \boldsymbol{U}_I^{\Gamma}(\mathcal{M}, \mathcal{W}, \succ)$.*

The proof is straightforward, but is provided in Appendix C.1 for completeness. In a biclique network the definitions of blocking–pairs and local blocking–pairs are equivalent, and all matchings trivially respect the network. Hence, the matching concepts trivially coincide under a biclique network.

The definition of technologically stable network matchings is more restrictive than the definition of stable matchings, in that weakly fewer matchings are considered feasible; however, it is less restrictive in that the set of potential blocking–pairs is weakly smaller. These forces suggest that the sets of stable matchings and technologically network–stable matchings cannot be generically nested; Proposition 4.3.5 confirms this intuition.

Proposition 4.3.5. *For any set of men $\mathcal{M}$, any set of women $\mathcal{W}$, and any preference profile $\succ \in \boldsymbol{P}$ such that there exist $m \in \mathcal{M}$ and $w \in \mathcal{W}$ with $w \succ_m m$ and $m \succ_w w$, there exists a network $\Gamma \subseteq \mathcal{M} \times \mathcal{W}$ such that $\boldsymbol{U}(\mathcal{M}, \mathcal{W}, \succ) \not\subseteq \boldsymbol{U}_T^{\Gamma}(\mathcal{M}, \mathcal{W}, \succ)$ and $\boldsymbol{U}_T^{\Gamma}(\mathcal{M}, \mathcal{W}, \succ) \not\subseteq \boldsymbol{U}(\mathcal{M}, \mathcal{W}, \succ)$.*

The proof is available in Appendix C.2. The restriction on the preferences is both necessary and sufficient for the result to hold; if all men find all women unacceptable at $\succ$ or all women find all men unacceptable $\succ$, then the only matching that is stable on either the marriage problem or any associated marriage network is that in which each player is unmatched. The following example illustrates the construction of such a network, given a suitable preference profile.

Example 4.3.6. *Let $\mathcal{N} = \{m_1, m_2\}$, $\mathcal{W} = \{w_1, w_2\}$, and let preferences $\succ$ be given by*

$$\succ = \begin{cases} m_1 : & w_1 \succ w_2 \succ m_1 \\ m_2 : & w_1 \succ w_2 \succ m_2 \\ w_1 : & m_1 \succ m_2 \succ w_1 \\ w_2 : & m_2 \succ m_1 \succ w_2. \end{cases} \tag{4.4}$$

In particular, every man finds every woman acceptable, and every woman finds every man acceptable. Consider the network given by $\Gamma = \{(m_1, w_2), (m_2, w_1), (m_2, w_2)\}$, depicted in Figure 4.1.

Consider the matching $\mu = \{(m_1, w_2), (m_2, w_1)\}$, and note that μ respects Γ, is individually rational, and has no local blocking pairs. Hence, $\mu \in \boldsymbol{U}_T^{\Gamma}(\mathcal{M}, \mathcal{W}, \succ)$. However, $\mu \notin \boldsymbol{U}(\mathcal{M}, \mathcal{W}, \succ)$, since (m_1, w_1) forms a blocking pair for μ in the marriage

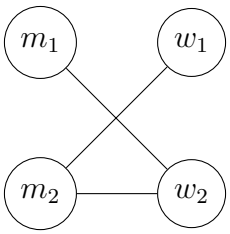


Figure 4.1: Network Γ, from Example 4.3.6.

problem $\langle \mathcal{M}, \mathcal{W}, \succ \rangle$.

Similarly, consider the matching $\nu = \{(m_1, w_1), (m_2, w_2)\}$*, and note that* ν *is individually rational and has no blocking pairs. Hence,* $\nu \in \boldsymbol{U}(\mathcal{M}, \mathcal{W}, \succ)$*. However,* $\nu \notin \boldsymbol{U}_T^{\Gamma}(\mathcal{M}, \mathcal{W}, \succ)$ *since it does not respect* Γ*. In particular,* $(m_1, w_1) \notin \Gamma$*.*

Therefore, Γ *constitutes a network for which* $\boldsymbol{U}(\mathcal{M}, \mathcal{W}, \succ) \not\subseteq \boldsymbol{U}_T^{\Gamma}(\mathcal{M}, \mathcal{W}, \succ)$ *and* $\boldsymbol{U}_T^{\Gamma}(\mathcal{M}, \mathcal{W}, \succ) \not\subseteq \boldsymbol{U}(\mathcal{M}, \mathcal{W}, \succ)$

Proposition 4.3.5 shows that the sets $\mathbf{U}(\mathcal{M}, \mathcal{W}, \succ)$ and $\mathbf{U}_T^{\Gamma}(\mathcal{M}, \mathcal{W}, \succ)$ cannot be nested in a way that is robust to all networks Γ for a suitable fixed preference profile $\succ$. Conversely, Proposition 4.3.7 shows that the sets cannot be nested in a way that is robust to all preference profiles $\succ$ for a suitable fixed network structure Γ.

Proposition 4.3.7. *For any set of men* $\mathcal{M}$*, any set of women* $\mathcal{W}$*, and any network structure* $\Gamma \subseteq \mathcal{M} \times \mathcal{W}$ *that is not biclique, there exists a preference profile* $\succ \in \boldsymbol{P}$ *such that* $\boldsymbol{U}(\mathcal{M}, \mathcal{W}, \succ) \not\subseteq \boldsymbol{U}_T^{\Gamma}(\mathcal{M}, \mathcal{W}, \succ)$ *and* $\boldsymbol{U}_T^{\Gamma}(\mathcal{M}, \mathcal{W}, \succ) \not\subseteq \boldsymbol{U}(\mathcal{M}, \mathcal{W}, \succ)$*.*

The proof is available in Appendix C.3, and proceeds by constructing a preference profile as a function of the given network Γ that yields the desired result. It should be noted that there may exist other preference profiles that lead to the same conclusion; in particular, Example 4.3.6 could be reinterpreted as a demonstration of

the construction of a preference profile, distinct from that specified in Appendix C.3, under which the sets $\mathbf{U}(\mathcal{M}, \mathcal{W}, \succ)$ and $\mathbf{U}_T^{\Gamma}(\mathcal{M}, \mathcal{W}, \succ)$ cannot be nested.

Lemma 4.3.4 and Propositions 4.3.5 and 4.3.7 are captured in Figure 4.2, where the size of the intersection of the set of stable matchings and the set of technologically network–stable matchings depends on the specific network and preference profile under consideration.

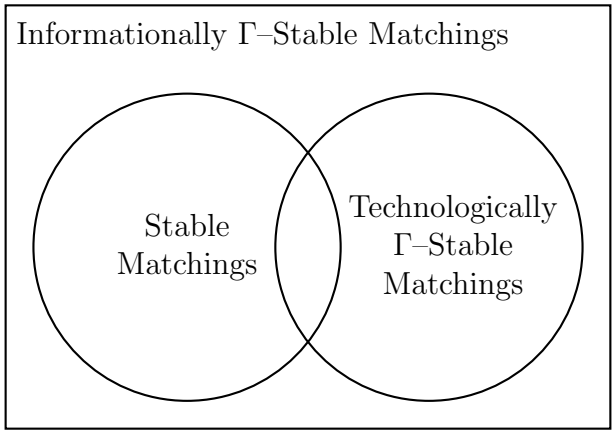


Figure 4.2: Generic Nesting Relationships of Stable Matching Concepts

Having analyzed the relationships between the various sets of stable matchings, Theorem 4.3.9 and its corollary establish the existence of informationally and technologically network–stable matchings.

Theorem 4.3.8. *A technologically* Γ*–stable matching exists for every marriage network* $\langle \mathcal{M}, \mathcal{W}, \succ, \Gamma \rangle$*; that is,* $U_T^{\Gamma}(\mathcal{M}, \mathcal{W}, \succ) \neq \emptyset$ *for all* $\langle \mathcal{M}, \mathcal{W}, \succ, \Gamma \rangle$*.*

The proof is available in Appendix C.4, in which a modified version of the Gale–Shapley Deferred Acceptance Algorithm (1962) is used to explicitly construct a technologically Γ–stable matching. Intuitively, if $(m, w) \notin \Gamma$, the modified algorithm treats man m as if he were unacceptable to woman w, and woman w as if she was

unacceptable to man m. See Appendix C.4 for a formal definition of the Network–Respecting Deferred Acceptance Algorithm (NDAA) and proof that for any marriage network $\langle \mathcal{M}, \mathcal{W}, \succ, \Gamma \rangle$, it is well–defined, terminates in finite time, and always produces a technologically Γ–stable matching.

Theorem 4.3.8 and Lemma 4.3.1 yield the following corollary, which can also be derived from Lemma 4.3.3 in the context of Gale & Shapley (1962).

Corollary 4.3.9. *An informationally Γ–stable matching exists for every marriage network $\langle \mathcal{M}, \mathcal{W}, \succ, \Gamma \rangle$; that is, $U_I^{\Gamma}(\mathcal{M}, \mathcal{W}, \succ) \neq \emptyset$ for all $\langle \mathcal{M}, \mathcal{W}, \succ, \Gamma \rangle$.*

Having established that informationally and technologically network–stable matchings exist, the next proposition justifies consideration of them from the perspective of Parteo optimality. Given a marriage problem $\langle \mathcal{M}, \mathcal{W}, \succ \rangle$, the set of stable matchings is Pareto undominated in the set of matchings. The proof is by contradiction: suppose that stable matching μ is Pareto dominated by matching ν, so that by definition $\nu \succsim_i \mu$ for all $i \in \mathcal{N}$ and there exists $j \in \mathcal{N}$ for whom $\nu \succ_j \mu$. Note in particular that this implies that $\nu(j) \neq \mu(j)$ and hence $\mu(\nu(j)) \neq j$. Since $\nu \succsim_{\nu(j)} \mu$, it therefore follows that $j \succ_{\nu(j)} \mu(\nu(j))$, so that the matching μ is blocked by $(j, \nu(j))$, a contradiction. An analogous result holds for the set of technologically Γ–stable matchings on the marriage network $\langle \mathcal{M}, \mathcal{W}, \succ, \Gamma \rangle$.

Proposition 4.3.10. *For any marriage network $\langle \mathcal{M}, \mathcal{W}, \succ, \Gamma \rangle$, the set of technologically Γ–stable matchings is Pareto dominant in the set of matchings that respect Γ.*

The proof is available in Appendix C.5. An analogous result does not hold for the set of informationally network–stable matchings, as the next proposition posits.

Proposition 4.3.11. *For any set of men $\mathcal{M}$, any set of women $\mathcal{W}$, and any preference profile $\succ \in \boldsymbol{P}$ such that there exist $m \in \mathcal{M}$ and $w \in \mathcal{W}$ with $w \succ_m m$ and $m \succ_w w$, there exists $\Gamma \subseteq \mathcal{M} \times \mathcal{W}$ such that the set of informationally Γ–stable matchings contains matchings that are Pareto dominated within $U_I^\Gamma(\mathcal{M}, \mathcal{W}, \succ)$.*

The proof is available in Appendix C.6, and proceeds by constructing a network Γ on which there is a Pareto dominated informationally Γ–stable matching. The result is not surprising: since players must be acquainted in order to form a local–blocking pair, it is easy to construct networks in which two unacquainted players could be made strictly better off by being matched to each other without detriment to the other players.

As in Proposition 4.3.5, the restriction on preferences in Proposition 4.3.11 is necessary for the construction to yield the negative result; if all men find all women unacceptable at $\succ$ or all women find all men unacceptable $\succ$, then the only matching that is informationally network–stable is that in which each player is unmatched. Under such preferences, this null matching is uniquely Pareto optimal in the set of all matchings.

It is an interesting and open question as to whether there exist general joint conditions on the preference profile $\succ$ and the network Γ that ensure that the set of informationally network–stable matchings on marriage network $\langle \mathcal{M}, \mathcal{W}, \succ, \Gamma \rangle$ is Pareto undominated. For the special case of a biclique network, Lemma 4.3.4 and Proposition 4.3.10 yield the following result.

Corollary 4.3.12. *For any set of men $\mathcal{M}$ and any set of women $\mathcal{W}$, if network $\Gamma \subseteq \mathcal{M} \times \mathcal{W}$ is biclique, then the set of informationally Γ–stable matchings is Pareto*

dominant in the set of matchings.

In addition to Pareto optimality, there are two analogous optimality conditions that can be naturally applied to marriage networks. The introduction of these conditions requires two new definitions.

Definition 4.3.13. *Given a marriage network $\langle \mathcal{M}, \mathcal{W}, \succ, \Gamma \rangle$, a technologically network–stable matching $\mu \in \boldsymbol{U}_T^\Gamma(\mathcal{M}, \mathcal{W}, \succ)$ is $\mathcal{M}_T$–**optimal** if $\mu \succsim_m \nu$ for all $m \in \mathcal{M}$ and all $\nu \in \boldsymbol{U}_T^\Gamma(\mathcal{M}, \mathcal{W}, \succ)$. Similarly, a technologically network–stable matching $\mu \in \boldsymbol{U}_T^\Gamma(\mathcal{M}, \mathcal{W}, \succ)$ is $\mathcal{W}_T$–**optimal** if $\mu \succsim_w \nu$ for all $w \in \mathcal{W}$ and all $\nu \in \boldsymbol{U}_T^\Gamma(\mathcal{M}, \mathcal{W}, \succ)$.*

Definition 4.3.14. *Given a marriage network $\langle \mathcal{M}, \mathcal{W}, \succ, \Gamma \rangle$, an informationally network–stable matching $\mu \in \boldsymbol{U}_I^\Gamma(\mathcal{M}, \mathcal{W}, \succ)$ is $\mathcal{M}_I$–**optimal** if $\mu \succsim_m \nu$ for all $m \in \mathcal{M}$ and all $\nu \in \boldsymbol{U}_I^\Gamma(\mathcal{M}, \mathcal{W}, \succ)$. Similarly, an informationally network–stable matching $\mu \in \boldsymbol{U}_I^\Gamma(\mathcal{M}, \mathcal{W}, \succ)$ is $\mathcal{W}_I$–**optimal** if $\mu \succsim_w \nu$ for all $w \in \mathcal{W}$ and all $\nu \in \boldsymbol{U}_I^\Gamma(\mathcal{M}, \mathcal{W}, \succ)$.*

The following theorem states that $\mathcal{M}_T$–optimal and $\mathcal{W}_T$–optimal technologically network–stable matchings exists for every marriage network $\langle \mathcal{M}, \mathcal{W}, \succ, \Gamma \rangle$, and provides an algorithm for computing them.

Theorem 4.3.15. *When all men and all women have strict preferences, there exists a $\mathcal{M}_T$–optimal technologically Γ–stable matching and a $\mathcal{W}_T$–optimal technologically Γ–stable matching for every marriage network $\Gamma \subseteq \mathcal{M} \times \mathcal{W}$. Furthermore, the matching $\mu_{\mathcal{M}_T}$ produced by the men–proposing network–respecting deferred acceptance algorithm is the unique $\mathcal{M}_T$–optimal technologically Γ–stable matching. Similarly, the unique $\mathcal{W}_T$–optimal technologically Γ–stable matching is produced by the women–proposing network–respecting deferred acceptance algorithm.*

The proof is available in Appendix C.7, and is parallel to that in Gale & Shapley (1962). On the set of informationally network–stable matchings, optimal stable matchings need not exist, as the following theorem states.

Theorem 4.3.16. *There exist marriage networks* $\langle \mathcal{M}, \mathcal{W}, \succ, \Gamma \rangle$ *for which no* $\mathcal{M}_I$*–optimal or* $\mathcal{W}_I$*–optimal informationally* Γ*–stable matchings exist.*

Example 4.3.17 illustrates the result of Theorem 4.3.16. It is an open question as to whether there exist conditions on marriage networks $\langle \mathcal{M}, \mathcal{W}, \succ, \Gamma \rangle$ under which the existence of $\mathcal{M}_I$–optimal and $\mathcal{W}_I$–optimal informationally Γ–stable matchings is guaranteed.

Example 4.3.17. *Let* $\mathcal{N} = \{m_1, m_2\}$*,* $\mathcal{W} = \{w_1, w_2\}$*, and* $\succ$ *be given by*

$$\succ = \begin{cases} m_1: & w_1 \succ w_2 \succ m_1 \\ m_2: & w_1 \succ w_2 \succ m_2 \\ w_1: & m_1 \succ m_2 \succ w_1 \\ w_2: & m_1 \succ m_2 \succ w_2. \end{cases} \tag{4.5}$$

Let the network structure be given by $\Gamma = \{(m_1, w_2), (m_2, w_1), (m_2, w_2)\}$*, as depicted in Figure 4.3.*

It can be shown that

$$\boldsymbol{U}_I^{\Gamma}(\mathcal{M}, \mathcal{W}, \succ) = \left\{ \begin{array}{rcl} \mu & = & \{(m_1, w_1), (m_2, w_2)\} \\ \nu & = & \{(m_1, w_2), (m_2, w_1)\} \end{array} \right\}. \tag{4.6}$$

Moreover, $\mu \succ_{m_1} \nu$ *but* $\nu \succ_{m_2} \mu$*, and* $\mu \succ_{w_1} \nu$ *but* $\nu \succ_{w_2} \mu$*. Hence, no* $\mathcal{M}_I$*–optimal*

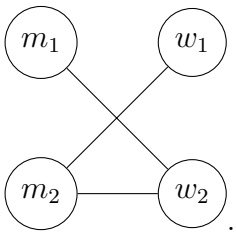


Figure 4.3: Network Γ, from Example 4.3.17

nor $\mathcal{W}_I$–optimal informationally Γ–stable matching exists for the given marriage network.

Extending our notation, for any two matchings μ and ν, let $\mu \succsim_{\mathcal{M}} \nu$ denote that $\mu \succsim_m \nu$ for all $m \in \mathcal{M}$; let $\mu \succsim_{\mathcal{W}} \nu$ be defined analogously.

Theorem 4.3.18. *Given a marriage network $\langle \mathcal{M}, \mathcal{W}, \succ, \Gamma \rangle$, the common preferences of the two sides of the markets are opposed on the set of technologically Γ–stable matchings. That is, if $\mu, \nu \in \boldsymbol{U}_T^{\Gamma}(\mathcal{M}, \mathcal{W}, \succ, \Gamma)$, then $\mu \succsim_{\mathcal{M}} \nu$ if and only if $\nu \succsim_{\mathcal{W}} \mu$.*

The proof is availble in Appendix C.8, and is parallel to that in Knuth (1976). The following corollary is immediate from Theorems 4.3.8 and 4.3.18.

Corollary 4.3.19. *Given a marriage network $\langle \mathcal{M}, \mathcal{W}, \succ, \Gamma \rangle$, the $\mathcal{M}_T$–optimal technologically Γ–stable matching is the worst technologically stable matching for the women. Similarly, the $\mathcal{W}_T$–optimal technologically Γ–stable matching matches each man with his least preferred achievable mate.*

Given two matchings μ and ν, define the "meet" of μ and ν as

$$(\mu \wedge \nu)(i) = \begin{cases} \min\limits_{\succ_i} \{\mu(i), \nu(i)\}, \text{ if } i \in \mathcal{M} \\ \max\limits_{\succ_i} \{\mu(i), \nu(i)\}, \text{ if } i \in \mathcal{W}. \end{cases} \tag{4.7}$$

That is, the meet of matchings μ and ν assigns to each man his least–preferred mate and to each woman her most–preferred mate from the two matchings. Similarly, we can define the "join" of μ and ν as

$$(\mu \vee \nu)(i) = \begin{cases} \max\limits_{\succ_i}\{\mu(i), \nu(i)\}, \text{ if } i \in \mathcal{M} \\ \min\limits_{\succ_i}\{\mu(i), \nu(i)\}, \text{ if } i \in \mathcal{W}, \end{cases} \tag{4.8}$$

which assigns to each man his most–preferred mate and to each woman her least–preferred mate from the two matchings.

Conway shows that the set of stable matches on the marriage problem $\langle \mathcal{M}, \mathcal{W}, \succ \rangle$ has a lattice structure. That is, if μ and ν are stable matchings, the resulting objects $(\mu \wedge \nu)$ and $(\mu \vee \nu)$ are also stable matchings. This result survives translation to the set of technologically network–stable matchings.

Theorem 4.3.20. *Given a marriage network* $\langle \mathcal{M}, \mathcal{W}, \succ, \Gamma \rangle$*, if* $\mu, \nu \in \boldsymbol{U}_T^{\Gamma}(\mathcal{M}, \mathcal{W}, \succ)$*, then* $(\mu \vee \nu), (\mu \wedge \nu) \in \boldsymbol{U}_T^{\Gamma}(\mathcal{M}, \mathcal{W}, \succ)$*.*

The proof is available in Appendix C.9. As the next example shows, the set of informationally network–stable matchings need not have a lattice structure.

Example 4.3.21. *Let* $\mathcal{N} = \{m_1, m_2\}$*,* $\mathcal{W} = \{w_1, w_2\}$*, and* $\succ$ *be given by*

$$\succ = \begin{cases} m_1: & w_1 \succ w_2 \succ m_1 \\ m_2: & w_1 \succ w_2 \succ m_2 \\ w_1: & m_1 \succ m_2 \succ w_1 \\ w_2: & m_1 \succ m_2 \succ w_2. \end{cases} \tag{4.9}$$

Let the network structure be given by $\Gamma = \{(m_1, w_2), (m_2, w_1), (m_2, w_2)\}$*, as depicted in Figure 4.4.*

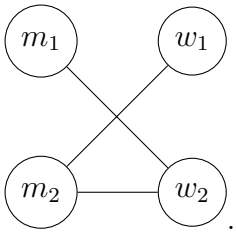


Figure 4.4: Network Γ, from Example 4.3.21

It can be easily verified that

$$\mu = \{(m_1, w_1), (m_2, w_2)\} \in \boldsymbol{U}_I^{\Gamma}(\mathcal{M}, \mathcal{W}, \succ) \tag{4.10}$$

and

$$\nu = \{(m_1, w_2), (m_2, w_1)\} \in \boldsymbol{U}_I^{\Gamma}(\mathcal{M}, \mathcal{W}, \succ). \tag{4.11}$$

However, the meet and join given by

$$(\mu \wedge \nu) = \{(m_1, w_2), (m_2, w_2), (m_1, w_1), (m_2, w_2)\} \tag{4.12}$$

and

$$(\mu \vee \nu) = \{(m_1, w_1), (m_2, w_1), (m_2, w_1), (m_1, w_2)\} \tag{4.13}$$

are not well–defined matchings and hence are not informationally network–stable matchings.

While Example 4.3.21 illustrates a particular marriage network in which the set

of informally network–stable matchings does not have lattice structure, it remains an open question whether there exist necessary and sufficient conditions on a marriage network under which the set of informally network–stable matchings does not have a lattice structure.

4.4 Strategic Network Formation

Having established the framework for analyzing networked marriage problems on fixed exogenously given networks, it is natural to embed such a model into a network formation game. This translation requires the construction of additional machinery and the donning of assumptions not typically made within the matching literature.

To adapt the marriage network framework to a game theoretic model, agents' ordinal utility must be translated to cardinal utility. For each man $m \in \mathcal{M}$, let $u_m(\cdot)$ be a utility function on $\mathcal{W}$ representing $\succ_m$; similarly, define $u_w(\cdot)$ for each woman. The appropriate functional forms of these utility functions remains an open modeling question, and will likely depend heavily on the context in which the model is being applied.

Each man $m \in \mathcal{M}$ is endowed with a pure strategy space $S_m = \{0,1\}^W$ with representative element $s_m(w) \in S_m$, where $s_m(w) = 1$ if man m proposes a link to woman w. Each woman has an analogous pure strategy space, and each agent has a mixed strategy space given by the unit simplex $\Delta(S_i)$ on S_i with representative element σ_i.

The network–formation technology and cost structure requires a number of additional modeling assumptions. In particular, what is required for a link to form

between man m and woman w? What is the cost of proposing a link, and is there an additional cost to link formation? For the purposes of this exercise, the network–formation technology is assumed to correspond to an "and" network: $(m, w) \in \Gamma$ if and only if $s_m(w) = 1$ and $s_w(m) = 1$. When players employ mixed strategies, the resulting network will depend on the specific realization of link proposals. Link formation is assumed to be costly, with each member of a link $(m, w) \in \Gamma$ incurring a cost c when the link is formed. To reduce the multiplicity of equilibria, link proposition is also assumed to costly, with each unrealized but proposed link costing the proposing agent ϵ. Under this modeling framework, when the pure strategy profile s is played, agent i pays a link formation cost of

$$c_i(s) = \sum_{j \in \mathcal{N} \setminus \{i\}} \mathbb{I}_{[s_i(j)=1]} \left(\epsilon + (c - \epsilon) \mathbb{I}_{[(i,j) \in \Gamma]} \right). \tag{4.14}$$

In the complete information setting, the network formation game proceeds as follows.

$t = 0$: A preference profile $\succ$ is realized and is common knowledge.

$t = 1$: Players simultaneously announce strategies σ_i.

$t = 2$: Network Γ is randomly drawn from the distribution on bipartite graphs produced by σ.

$t = 3$: A network–stable matching μ is chosen.

$t = 4$: Each player $i \in \mathcal{N}$ realizes utility given by

$$u_i(\mu(i)) - \sum_{(i,j)\in\Gamma} c - \sum_{(i,j)\notin\Gamma} \epsilon\sigma_i(j). \tag{4.15}$$

It remains an important modeling question how the network–stable matching should be selected when the sets $\mathbf{U}_I^\Gamma(\mathcal{M}, \mathcal{W}, \succ)$ and $\mathbf{U}_T^\Gamma(\mathcal{M}, \mathcal{W}, \succ)$ are not singleton.

In this framework, the natural definition of a Nash equilibrium is given as follows.

Definition 4.4.1. *A mixed–strategy profile σ is a **pairwise–Nash equilibrium** if for all pairs $(m, w) \in \mathcal{M} \times \mathcal{W}$ and all mixed strategies $\tilde{\sigma}_m$, $\tilde{\sigma}_w$,*

$$\mathbb{E}[u_m(\sigma_{-(m,w)}, \tilde{\sigma}_m, \tilde{\sigma}_w)] < \mathbb{E}[u_m(\sigma)]. \tag{4.16}$$

Note that in Definition 4.4.1, the expectation operator is taken both with respect to the distribution on bipartite networks produced by the mixed strategy profiles as well as the distribution over network–stable matchings determined by the mechanism. If it is assumed that a network–stable matching is selected uniformly at random from $\mathbf{U}_T^\Gamma(\mathcal{M}, \mathcal{W}, \succ)$, the following result obtains in the complete information case.

Theorem 4.4.2. *Consider a marriage network formation game $\langle \mathcal{M}, \mathcal{W}, \{u_i\}_{i\in\mathcal{N}} \rangle$ in which preferences are common knowledge. Suppose that there is a link–proposal cost of $\epsilon > 0$ and a link–formation cost of c. If $\epsilon < c$ and*

$$c < \min_{(i,j)\ s.t.\ \exists \mu \in \boldsymbol{U}(\mathcal{M},\mathcal{W},\succ)\ s.t.\ \mu(i)=j} \{u_i(j) - u_i(i)\},$$

then the pure–strategy profile $s^\star$ is a pairwise–Nash equilibrium if and only if there

exists a technologically Γ*a–stable matching* μ *on the marriage problem* $\langle \mathcal{M}, \mathcal{W}, \succ \rangle$ *such that*

$$s_i^\star(j) = \begin{cases} 1, & \text{if } \mu(i) = j \\ 0, & \text{otherwise.} \end{cases} \tag{4.17}$$

The proof of Theorem 4.4.2 is available in Appendix C.10, and appeals to the following lemma.

Lemma 4.4.3. *Consider a marriage network formation game* $\langle \mathcal{M}, \mathcal{W}, \{u_i\}_{i \in \mathcal{N}} \rangle$*. Suppose that there is a link–proposal cost of* $\epsilon > 0$ *and a link–formation cost of* c*. If* $s^\star$ *is a pure–strategy pairwise–Nash equilibrium, then* $\boldsymbol{U}_T^{\Gamma(s^\star)}$ *is a singleton. Moreover,* $s_i^\star(j) = 1$ *if and only if* $\mu(i) = j$*, where* $\boldsymbol{U}_T^{\Gamma(s^\star)} = \{\mu\}$*.*

The proof of Lemma 4.4.3 is available in Appendix C.11. Theorem 4.4.2 should be understood as a formal confirmation that embedding the marriage problem with a network structure does not fundamentally alter the model under full information assumptions; that is, network–respecting stability concepts have no bite in a full information setting. The true test of the theory of marriage networks lies in the equilibrium predictions of network formation models in which preferences are private information. These predictions remain an open question, and solving for the set of pairwise–Nash equilibria in models in which preferences are privately known and uncorrelated presently appears intractable.

4.5 Conclusion

In large markets, the implicit assumption of common knowledge of preferences inherent to the matching literature may lead to predictions that differ substantially from observed market outcomes. By embedding the standard matching framework into a social network, the informational limitations of the agents can be explicitly captured in a structure amenable to economic and mathematical analysis. The framework and results presented here set the stage for future researchers. In particular, the question of characterizing the set of pairwise–Nash equilibria under informational settings other than common knowledge of preferences remains incredibly important and open.

This paper presents a particular model of networked one–to–one matching, and establishes a number of properties of such markets analogous to those presented in Roth and Sotomayor (1962) on the standard marriage problem. Where analogous results cannot be obtained, a number of examples are presented illustrating the mechanism through which certain properties fail to hold. Finally, a bare–bones network formation game is considered and it is shown that under full information, embedding a marriage problem with a network structure does not meaningfully alter the set of equilibrium predictions.

The true test of the marriage network framework lies in equilibrium predictions for network formation games that precede a networked matching procedure. Theorem 4.4.2 characterizes the set of pure–strategy pairwise–Nash equilibria under full information, and verifies that the additional machinery of the marriage network framework has no effect on the set of equilibria in the absence of informational restrictions. It remains to determine what equilibrium characterizations are possible under more

restrictive assumptions on agents' information sets. As is endemic within the literature on social networks, model tractability and the absence of closed–form analytic solutions remain the most considerable barriers to further characterization results. Barring technical or mathematical innovations that expand the scope of analytically tractable network questions, employing alternative modeling assumptions may allow for incremental progress. The network formation technology – that is, under what link proposal conditions network links are formed – and the cost structure of link proposal and formation present themselves as the modeling assumptions most ripe for reconsideration.

While not considered in this paper, adapting the model to allow for informational exchange between agents who are not directly connected in a network would expand the scope of the model and align it more closely with the social networks literature. Following Jackson and Wolinsky (1996), Dutta and Mutuswami (1997), Bala and Goyal (1999), and Jackson and Rogers (2005), a candidate model of this type would allow for agent i to learn about the existence and preferences of agent j with a probability that decreases exponentially in the length of the shortest path connecting i to j in Γ. It is worth noting that the incentives for link formation would differ markedly in such a model from those in the model presented here; in particular, in a model in which agents can learn about the existence and preferences of agents with whom they are only indirectly connected, agents have an incentive to form within–group links in addition to between–group links.

Bibliography

[1] Abdulkadiroğlu, A. and T. Sönmez (2003): "School Choice: A Mechanism Design Approach," *American Economic Review*, 93(3), 729–747.

[2] Abreu, D. and M. Manea (2009a): "Bargaining and Efficiency in Networks," mimeo.

[3] Abreu, D. and M. Manea (2009b): "Markov Perfect Equilibria in a Model of Bargaining in Networks," mimeo.

[4] Ali, S. N. (2011): "Learning Self–Control," *The Quarterly Journal of Economics*, 126, 857–893.

[5] Arcaute, E. and S. Vassilvitskii (2009): "Social Networks and Stable Matchings in the Job Market," mimeo.

[6] Arrow, K. J. (1951): *Social Choice and Individual Values*. New York: John Wiley and Sons, Inc.

[7] Arcaute, E. and S. Vassilvitskii (2009): "Social Networks and Stable Matchings in the Job Market." The Workshop on Internet & Network Economics, Sapienza University of Rome, Italy.

[8] Asheim, G. B. (2007): "Procrastination, Partial Naivete, and Behavioral Welfare Analysis," mimeo.

[9] Aswal, N., S. Chatterji, and A. Sen (2003): "Dictatorial Domains," *Economic Theory*, 22, 45–62.

[10] Bala, V. and S. Goyal (2000): "A Non–Cooperative Model of Network Formation," *Econometrica*, 68(5), 1181–1229.

[11] Barro, R. (1999): "Ramsey Meets Laibson in The Neoclassical Growth Model," *Quarterly Journal of Economics*, 114, 1125–1152.

[12] Bénabou, R. and J. Tirole (2004): "Willpower and Personal Rules," *Journal of Political Economy*, 112, 848–886.

[13] Billingsly, P. (1968): *Convergence of Probability Measures*. New York: John Wiley and Sons, Inc.

[14] Bloch, F. and M.O. Jackson (2007): "The Formation of Networks with Transfers Among Players," *Journal of Economic Theory*, 133(1), 83–110.

[15] Bodner, R. and D. Prelec (1997): "The Diagnostic Value of Actions in a Self–Signaling Model," mimeo.

[16] Bodner, R. and D. Prelec (2002): "Self–Signaling in a Neo-Calvinist Model of Everyday Decision Making," *Psychology and Economics*, 1.

[17] Campbell, D. and J. Kelly (2000): "A Simple Characterization of Majority Rule," *Economic Theory*, 15, 689–700.

[18] Campbell, D. and J. Kelly (2003): "A Strategy–Proofness Characterization of Majority Rule," *Economic Theory*, 22, 557–568.

[19] Campbell, D. and J. Kelly (2006): "Social Welfare Functions Generating Social Choice Rules That Are Invulnerable to Manipulation," *Mathematical Social Sciences*, 51, 81–89.

[20] Corominas–Bosch, M. (2004): "Bargaining in a Network of Buyers and Sellers," *Journal Economic Theory*, 115, 35–77.

[21] Della Vigna, S. and U. Malmendier (2006): "Paying Not to Go to the Gym," *American Economic Review*, 96(3), 694–719.

[22] Doob, J. L. (1953): *Stochastic Processes*. New York: John Wiley and Sons, Inc.

[23] Dutta, B. and S. Mutuswami (1997): "Stable Networks," *Journal of Economic Theory*, 76, 322–344.

[24] Eliaz, K. (2004): "Social Aggregators," *Social Choice and Welfare*, 22, 317–330.

[25] Eliaz, K. and R. Spiegler (2006): "Contracting with Diversely Naïve Agents," *Review of Economic Studies*, 73(3), 689–714.

[26] Fabrikant, A., A. Luthra, E. Maneva, C. Papadimitriou, and S. Shenker (2003): "On a Network Creation Game," Proceedings of ACM Symposium on Principles of Distributed Systems, New York: ACM.

[27] Fainmesser, I. (2009): "Community Structure and Market Outcomes: Towards a Theory of Repeated Games in Networks." Ph.D. Dissertation, Harvard University, Department of Economics.

[28] Fainmesser, I. and D. Goldberg (2009): "Effective Word–Of–Mouth: Reputation Networks and Market Structure." Ph.D. Dissertation, Harvard University, Department of Economics.

[29] Feldman, M. (1991): "On the Generic Nonconvergence of Bayesian Actions and Beliefs," *Economic Theory*, 1, 301–321.

[30] Fudenberg, D. (2006): "Advancing Beyond 'Advances in Behavioral Economics,"' *Journal of Economic Literature*, 44, 694–711.

[31] Fudenberg, D. and D. K. Levine (2006): "A Dual Self Model of Impulse Control," *American Economic Review*, 96, 1449–1476.

[32] Gale, D. and L. Shapley (1962): "College Admissions and the Stability of Marriage," *American Mathematical Monthly*, 69, 9–15.

[33] Gale, D. and M. Sotomayor (1985): "Some Remarks on the Stable Matching Problem," *Discrete Applied Mathematics*, 11(3), 223–232.

[34] Gibbard, A. (1973): "Manipulation of Voting Schemes: A General Result," *Econometrica*, 41, 587–601.

[35] Gul, F. and E. Stachetti (1999): "Walrasian Equilibrium with Gross Substitutes," *Journal of Economic Theory*, 87, 95–124.

[36] Heidhues, P. and B. Köszegi (2009): "Futile Attempts at Self–Control," *Journal of the European Economic Association*, y(2–3), 423–434.

[37] Jackson, M. O. (2008): *Social and Economic Networks*. Princeton: Princeton University Press.

[38] Jackson, M. O. and A. Wolinsky (1996): "A Strategic Model of Social and Economic Networks," *Journal of Economic Theory*, 71(1), 44–74.

[39] Jackson, M.O. and B.W. Rogers (2003): "The Economics of Small Worlds," *Journal of the European Economic Association (Papers and Proceedings)*, 3(2–3), 617–627.

[40] Kelso, A. S. and V. Crawford (1982): "Job Matching, Coalition Formation, and Gross Substitutes," *Econometrica*, 50, 1483–1504.

[41] Knuth, D. E. (1976): *Marriage Stables et leurs Relations avec d'autres Problèmes Combinatoires*. Montréal: Les Presses de l'Université de Montréal.

[42] Kojima, K. and P. Pathak (2009): "Incentives and Stability in Large Two–Sided Matching Markets, " *American Economic Review*, 99, 608–627.

[43] Kranton, R. and D. Minehart (2001): "A Theory of Buyer–Seller Networks," *American Economic Review*, 91, 485–508.

[44] Laibson, D. (1994): "Essays in Hyperbolic Discounting," Ph.D. Dissertation, Massachusetts Institute of Technology, Department of Economics.

[45] Laibson, D. (1995): "Hyperbolic Discount Functions, Undersaving, and Savings Policy," mimeo.

[46] Laibson, D. (1997): "Hyperbolic Discounting and Golden Eggs," *Quarterly Journal of Economics*, 112, 443–477.

[47] Liu, Q., G. J. Mailath, A. Postlewaite, and L. Samuelson (2012): "Matching with Incomplete Information," working paper forthcoming.

[48] Maskin, E. (1995): "Majority Rule, Social Welfare Functions, and Game Forms." In: Basu, K., P. K. Pattanaik, and K. Suzumura (editors), *Choice, Welfare, and Development*. Oxford: The Clarendon Press.

[49] May, K. (1952): "A Set of Independent Necessary and Sufficient Conditions for Simple Majority Decision," *Econometrica*, 20, 680–684.

[50] McVitie, D. and L. Wilson (1970): "Stable Marriage Assignment for Unequal Sets." *Behaviour & Information Technology*, 10, 295–309.

[51] Niederle, M. and A. E. Roth (2003): "Relationship Between Wages and Presence of a Matching in Medical Fellowships," *Journal of the American Medical Associations*, 290(9), 1153–1154.

[52] O'Donoghue, T. and M. Rabin (1999): "Doing It Now or Later," *American Economic Review* 89, 103–124.

[53] O'Donoghue, T. and M. Rabin (2001): "Choice and Procrastination," *Quarterly Journal of Economics*, 116, 121–160.

[54] Ostrovsky, M. (2008): "Stability in Supply Chain Networks," *American Economic Review*, 98(3), 897–923.

[55] Phelps, E. S. and R. A. Pollak (1968): "On Second–Best National Saving and Game–Equilibrium Growth," *Review of Economic Studies*, 35, 185–199.

[56] Polanski, A. (2007): "Bilateral Bargaining in Networks," *Journal Economic Theory*, 134(1), 557–565.

[57] Pollak, R. A. (1968): "Consistent Planning," *Review of Economic Studies*, 35, 201–208.

[58] Roth, A. E. (1982): "Incentive Compatibility in a Market with Indivisibilities," *Economic Letters*, 9, 127–132.

[59] Roth, A. E. (1984): "The Evolution of the Labor Market for Medical Interns and Residents: A Case Study in Game Theory," *Journal of Political Economy*, 92, 991–1016.

[60] Roth, A. E. and M. A. O. Sotomayor (1990): *Two–Sided Matching: A Study in Game–Theoretic Modeling and Analysis*. Cambridge: Cambridge University Press.

[61] Roth, A. E., T. Sönmez, and M. U. Ünver (2004): "Kidney Exchange," *Quarterly Journal of Economics*, 119(2), 457–488.

[62] Satterthwaite, M. (1975): "Strategy–Proofness and Arrow's Conditions: Existence and Correspondence Theorems for Voting Procedures and Social Welfare Functions," *Journal of Economic Theory*, 10, 187–217.

[63] Sönmez, T. (1997): "Manipulation via Capacities in Two–Sided Matching Markets," *Journal of Economic Theory*, 77, 197–204.

[64] Sönmez, T. (1999): "Can Pre–Arranged Matched be Avoided in Two-Sided Matching Markets?," *Journal of Economic Theory*, 86, 148–156.

[65] Watts, A. (2001): "A Dynamic Model of Network Formation," *Games and Economic Behavior*, 34, 331-341.

Appendix A

Appendix to Chapter 2

A.1 Supplemental Notation & Definitions

The following two definitions are used throughout Appendix A in proving the results of Chapter 2.

Given a preference profile p and alternatives $x, y \in X$, let $N_p(x,y) = \{i \in N : x \succ_i^p y\}$ denote the *coalition over* (x,y), and note that for all p, x, and y: (i) $N_p(x,y) \subseteq N$; (ii) $\#N_p(x,y) \in \mathbb{N}_+$; (iii) $N_p(x,y) \cap N_p(y,x) = \emptyset$, and; (iv) $N_p(x,y) \cup N_p(y,x) = N$. A coalition is a *majority coalition for x over y* if $\#N_p(x,y) > \frac{n}{2}$. Having introduced the notion of majority coalitions, note that equivalent definition of a strong Condorcet winner: alternative $x \in X$ is the strong Condorcet winner at profile $p \in A(X)^n$ if and only if $\#N_p(x,y) > \frac{n}{2}$ for all $y \in X \setminus \{x\}$.

Given profiles $p, q \in \wp$, the *standard sequence from p to q* is the set $\{q^t\}_{t=0}^n$, where $q^0 = p$ and for $t > 0$, $q^t(i) = q(i)$ for all $i \in \{1, \ldots, t\}$ and $q^t(i) = p(i)$ for all $i \in \{t+1, \ldots, n\}$. The standard sequence $\{q^t\}_{t=0}^n$ can be thought of as a "path"

consisting of $n+1$ profiles that starts at p and ends at q. It should be noted that for an arbitrary domain $\wp$, $p, q \in \wp$ does not imply that $\{q^t\}_{t=0}^n \subset \wp$.

A.2 Proof of Lemma 2.4.1

Proof. Let $p \in \wp_C$, so that by definition there exists a strong Condorcet winner $x \in X$ at preference profile p. Therefore, $p \in C_x \subset \cup_{x \in X} C_x$ and hence $\wp_C \subseteq \cup_{x \in X} C_x$.

Next, let $p \in \cup_{x \in X} C_x$. Then there exists $x \in X$ such that $p \in C_x$, so that by construction x is a strong Condorcet winner at profile p. Therefore $p \in \wp_C$ by definition, and hence $\cup_{x \in X} C_x \subseteq \wp_C$. Therefore, $\wp_C = \cup_{x \in X} C_x$.

Finally, suppose that $p \in C_x$ for some $x \in X$. Then by construction x is a strong Condorcet winner at profile p, so that for all $y \in X \setminus \{x\}$, $\#N_p(x, y) > \frac{n}{2}$ implies that $\#N_p(y, x) < \frac{n}{2}$, since the individual preference orderings in p are asymmetric. Therefore, y cannot be a strong Condorcet winner at p and hence $p \notin C_y$ for all $y \in X \setminus \{x\}$. That is, $C_x \cap C_y = \emptyset$ for all $x, y \in X$. □

The final step in the proof illustrates that if a strong Condorcet winner exists at profile p, it must be unique. This fact was observed in Section 2.2, but is made rigorous by appealing to the majority coalitions.

A.3 Proof of Lemma 2.4.2

Proof. By definition, $x \in X$ is the strong Condorcet winner at profile $p \in C_x$, so that for all $y \in X \setminus \{x\}$, $\#N_p(x, y) > \frac{n}{2}$. Since n is even and $\#N_p(x, y) \in \mathbb{N}_+$ all $x, y \in X$, the strict inequality can be expressed as the weak inequality $\#N_p(x, y) \geq \frac{n}{2} + 1$.

For $i \in N$, let $q \in \wp_C$ be such that $q(j) = p(j)$ for all $j \in N \setminus \{i\}$, and choose an arbitrary $y \in X \setminus \{x\}$.

If $x \succ_i^q y$, then $\#N_q(x,y) \geq \#N_p(x,y) > \frac{n}{2}$, so that $x \in X$ must be the strong Condorcet winner at profile q. If $y \succ_i^q x$, then $\#N_q(x,y) \geq \#N_p(x,y) - 1 \geq \frac{n}{2}$. If either inequality is strict, then $x \in X$ must be the strong Condorcet winner at profile q as above.

If both inequalities bind, then $\#N_q(x,y) = \frac{n}{2}$, so that neither alternatives $x, y \in X$ can be the strong Condorcet winner at q. Since $y \in X \setminus \{x\}$ was chosen arbitrarily, this implies that there does not exist a Condorcet winner at profile q, so that $q \notin \wp_C$, a contradiction.

Therefore, alternative x is the strong Condorcet winner at q, so that $q \in \mathcal{C}_x$. □

A.4 Proof of Lemma 2.4.3

Proof. The proof of necessity is straightforward: if g is strategy–proof over the entire domain $\wp_C$ it must also be strategy–proof over any subdomain. In particular, it must be strategy–proof over each Condorcet section $C_x \subset \wp_C$ for all $x \in X$.

To prove sufficiency, suppose that the restriction $g|_{C_x}$ is strategy–proof for each $x \in X$ but that g is not strategy–proof on $\wp_C$. Then there exist profiles $p, q \in \wp_C$ and an individual $i \in N$ such that $p(j) = q(j)$ for all $j \in N \setminus \{i\}$ and $g(q) \succ_p^i g(p)$.

Since the set of Condorcet sections produces a partition of $\wp_C$ by Lemma 2.4.1, there exists $x \in X$ such that $p \in C_x$. Moreover, $q \in C_x$ by Lemma 2.4.2. Then individual i can manipulate $g|_{C_x}$ at profile p via $q(i)$, a contradiction. Therefore, if $g|_{C_x}$ is strategy–proof for each $x \in X$, then g is strategy–proof on $\wp_C$. □

A.5 Proof of Proposition 2.4.5

Proof. The proof is immediate: if $\#g(C_x) = 1$, then for all $i \in N$, $p \in C_x$, and $q \in C_x$ such that $q(j) = p(j)$ for all $j \in N \setminus \{i\}$, $g|_{C_x}(p) = g|_{C_x}(q)$. Hence, no individual can manipulate and $g|_{C_x}$ is strategy–proof. □

Intuitively, since the restriction $g|_{C_x}$ is not responsive to preferences, no individual can manipulate $g|_{C_x}$ over its domain C_x. Note that this result is independent of whether g satisfies unanimity, in which case $g(C_x) = \{x\}$ necessarily. Note further that majority rule is such that $\#g(C_x) = 1$ for all $x \in X$, with $g(C_x) = \{x\}$, and is therefore strategy–proof on $\wp_C$.

A.6 Proof of Proposition 2.4.6

Proof. Let $x \in X$ be such that $\#g(C_x) = 2$, and let $g(C_x) = \{y, z\} \subset X$ (where possibly $z = z$ or $y = x$, but not both).

To show necessity, suppose that $g|_{C_x}$ does not satisfy non–reversal. Then there exist profiles $p, q \in C_x$ such that $g|_{C_x}(p) = z$ and $g|_{C_x}(q) = y$, and some $i \in N$ such that $y \succ_i^p z$, and $q(j) = p(j)$ for all $j \in N \setminus \{i\}$. Therefore, individual i could manipulate $g|_{C_x}$ at profile p via $q(i)$, so that $g|_{C_x}$ is not strategy–proof.

To show sufficiency, suppose that $g|_{C_x}$ satisfies non–reversal. For $p \in C_x$, if $g|_{C_x}(p) = z$ and $z \succ_i^p y$ for some $i \in N$, then individual i cannot precipitate the selection of a more preferred alternative by reporting a preference ordering other than $p(i)$, since $g(C_x) = \{y, z\}$. If $g|_{C_x}(p) = z$ and $y \succ_i^p z$ for some $i \in N$, then by the non–reversal condition individual i cannot precipitate the selection of y. Thus,

$g|_{C_x}$ is strategy–proof. □

Note that this result is independent of whether g satisfies unanimity, in which case $x \in g(C_x)$ necessarily. Furthermore, this result applies on more general domains: non–reversal is equivalent to strategy–proofness over any domain with a two–element range.

A.7 Proof of Proposition 2.4.7

Proof. Sufficiency is clear, since dictatorial rule is strategy–proof.

The proof of necessity consists of two cases: when $x \in g(C_x)$ and when $x \notin g(C_x)$. When $x \in g(C_x)$, the structure and content of the proof is similar to that of Campbell and Kelly (2003), but differs nontrivially at several crucial steps as a result of the structure of $\wp_C$ when n is even.

When $x \notin g(C_x)$, the proof consists of two steps. In the first step the Gibbard–Satterthwaite theorem (1973, 1975) is invoked over a subset $U_x \subset C_x$ to show that if $g|_{C_x}$ is strategy–proof on U_x, it must be dictatorial on U_x with respect to the alternatives in $g(C_x)$. In the second step, strategy–proofness is shown to imply that the dictator over U_x must be the dictator over the entirety of C_x.

Part 1. Suppose that $g|_{C_x}$ is strategy–proof, with $\#g(C_x) \geq 3$ and $x \in g(C_x)$.

Step 1.1: If $x \in g(C_x)$, then $g|_{C_x}(u) = x$ at every unanimous profile $u \in C_x$.

Let $p \in C_x$ be such that $g|_{C_x}(p) = x$, and let $u \in C_x$ be a unanimous profile at which $u_1(i) = x$ for all $i \in N$. Let $\{u^t\}_{t=1}^n$ be the standard sequence from profile p to u. Note that for all $t \in \{0, \ldots, n\}$, $\#N_{u^t}(x, y) \geq \#N_p(x, y) > \frac{n}{2}$, since $p \in C_x$ and

at each element u^t in the standard sequence, alternative x is promoted to maximal element of individual i's preference ordering. Therefore, $\{u^t\}_{t=1}^n \subset C_x$.

By assumption, $g|_{C_x}(p) \equiv g|_{C_x}(u^0) = x$. For $t \in \{0, \ldots, n\}$, suppose that $g|_{C_x}(u^t) = x$ but $g|_{C_x}(u^{t+1}) = y$ for some $y \in g(C_x) \setminus \{x\}$. Since $u_1^{t+1}(t+1) = x$, $x \succ_{t+1}^{u^{t+1}} y$ so that individual $t+1$ can manipulate $g|_{C_x}$ at u^{t+1} via $u^t(t+1)$, a contradiction. Therefore, it must be that $g|_{C_x}(u^{t+1}) = x$. Continuing the inductive argument, $g|_{C_x}(u) \equiv g|_{C_x}(u^n) = x$, so that $x \in g(C_x)$ implies that $g|_{C_x}(u) = x$ at every unanimous profile $u \in C_x$.

Step 1.2: If $g|_{C_x}$ is non–dictatorial, then $g|_{C_x}(p) = x$ for all profiles $p \in C_x$ for which $\#\{i \in N : p_1(i) = x\} > \frac{n}{2}$.

Define $K_x(p) = \{i \in N : p_1(i) = x\}$ as the set of individuals for whom $x \in X$ is maximal at profile p. From Step 1.1, note that $g|_{C_x}(p) = x$ whenever $\#K_x(p) = n$. For $\kappa \in \{\frac{n}{2} + 2, \ldots, n\}$, suppose that $g|_{C_x}(p) = x$ whenever $\#K_x(p) = \kappa$ and consider a profile $p \in C_x$ at which $\#K_x(p) = \kappa - 1$.

Suppose that $g|_{C_x}(p) = y$ for some $y \in g(C_x) \setminus \{x\}$. If $x \succ_i^p y$ for some $i \in N \setminus K_x(p)$, then the induction hypothesis implies that individual i could manipulate $g|_{C_x}$ at profile p via a preference ordering $q(i)$ for which $q_1(i) = x$, a contradiction. Therefore, $y \succ_i^p x$ for all $i \in N \setminus K_x(p)$.

Let profile q be such that $q(i) = p(i)$ for all $i \in K_x(p)$ and $q(i) = (y \succ x \succ \ldots)$ for all $i \in N \setminus K_x(p)$. Since $\#K_x(q) = \#K_x(p) > \frac{n}{2}$, $\#N_q(x, z) \geq \#K_x(q) > \frac{n}{2}$ for all $z \in g(C_x) \setminus \{x\}$, so that $q \in C_x$. Let $\{q^t\}_{t=0}^n$ be the standard sequence from p to q, and note that for all $z \in g(C_x) \setminus \{x\}$, $\#N_{q^t}(x, z) \geq \#K_x(q^t) = \#K_x(q) > \frac{n}{2}$, so that $\{q^t\}_{t=0}^n \subset C_x$.

By assumption, $g|_{C_x}(q^0) \equiv g|_{C_x}(p) = y$. For $t \in \{0, \ldots, n\}$, if $g|_{C_x}(q^t) = y$ and $g|_{C_x}(q^{t+1}) \neq y$, then necessarily $q^{t+1} \neq q^t$, which implies that $t+1 \in N \setminus K_x(p)$, $y \succ^q_{t+1} g|_{C_x}(q^{t+1})$, and hence $t+1$ can manipulate $g|_{C_x}$ at q^{t+1} via $q^t(t+1)$. Therefore, $g|_{C_x}(q^t) = y$ implies that $g|_{C_x}(q^{t+1}) = y$, so that $g|_{C_x}(q) \equiv g|_{C_x}(q^n) = y$ by induction.

Let profile r be such that for all $i \in K_x(p)$, $r_1(i) = x$ and $z \succ^r_i y$ for all $z \in g(C_x) \setminus \{y\}$, and $r(i) = q(i)$ for $i \in N \setminus K_x(p)$. Since $\#N_r(x,z) \geq \#K_x(r) = \#K_x(q) > \frac{n}{2}$ for all $z \in g(C_x) \setminus \{x\}$, $r \in C_x$. Let $\{r^t\}_{t=0}^n$ be the standard sequence from profile q to r, and note that $\#N_{r^t}(x,z) \geq \#K_x(r^t) = \#K_x(r) > \frac{n}{2}$, so that $\{r^t\}_{t=0}^n \subset C_x$. From above, $g|_{C_x}(r^0) \equiv g|_{C_x}(q) = y$.

For $t \in \{0, \ldots, n\}$, suppose that $g|_{C_x}(r^t) \in g(C_x) \setminus \{x, y\}$. Then for every $i \in N \setminus K_x(p)$, $x \succ^{r^t}_i g|_{C_x}(r^t)$, so that individual i can manipulate $g|_{C_x}$ at r^t via profile s for which $s(j) = r^t(j)$ for all $j \in N \setminus \{i\}$ and $s_1(i) = x$, since $K_x(s) = K_x(p) + 1 = \kappa$, by the induction hypothesis. Therefore, $g|_{C_x}(r^t) \in \{x, y\}$ for all $t \in \{0, \ldots, n\}$.

Suppose that for some $t \in \{0, \ldots, n\}$, $g|_{C_x}(r^t) = y$. If $g|_{C_x}(r^{t+1}) = x$, then $r^t \neq r^{t+1}$, which implies that $t+1 \in K_x(p)$ and hence $x \succ^{r^t}_{t+1} y$. Therefore, individual $t+1$ can manipulate $g|_{C_x}$ at r^t via $r^{t+1}(t+1)$, a contradiction. Therefore $g|_{C_x}(r^{t+1}) = y$, and by induction $g|_{C_x}(r) \equiv g|_{C_x}(r^n) = y$.

Denote by $C^\star_x$ the set of all profiles s such that $s(i) = r(i)$ for all $i \in K_x(p)$ and $s(i)$ is an arbitrary linear ordering on X for all $i \in N \setminus K_x(p)$, so that in particular $r \in C^\star_x$. Again, note that $\#N_s(x,z) \geq \#K_x(s) = \#K_x(r) > \frac{n}{2}$ for all $z \in g(C_x) \setminus \{x\}$, so that $C^\star_x \subset C_x$. Define a new social choice rule $g^\star : C^\star_x \to X$ by $g^\star(s) = g|_{C_x}(s)$ for all $s \in C^\star_x$.

Note that $r \in C^\star_x$, so that $g^\star(r) = y$ by construction and $y \in g^\star(C^\star_x)$. Next, fix an

individual $i \in N \setminus K_x(p)$ and consider a preference profile s such that $s_1(i) = x$ and $s(j) = r(j)$ for all $j \in N \setminus \{i\}$. Note that $\#K_x(s) = \#K_x(r) + 1 = \kappa$, so that $s \in C_x^\star$ and by the induction hypothesis $g^\star(s) = g|_{C_x}(s) = x$. Hence, $x \in g^\star(C_x^\star)$.

Let $z \in g(C_x) \setminus \{x, y\}$ and define s such that $s(i) = r(i)$ for all $i \in K_x(p)$ and $s(i) = (z \succ y \succ \dots)$ for all $i \in N \setminus K_x(p)$. Since $\#N_s(x, z) \geq \#K_x(s) = \#K_x(r) > \frac{n}{2}$ for all $z \in g(C_x) \setminus \{x\}$, $s \in C_x^\star$ by construction. Let $\{s^t\}_{t=0}^n$ be the standard sequence from r to s. Note that for all $t \in \{0, \dots, n\}$, $\#N_{s^t}(x, z) \geq \#K_x(s^t) = \#K_x(s) > \frac{n}{2}$, so that $\{s^t\}_{t=0}^n \subset C_x^\star$. By definition, $g^\star(s^0) \equiv g^\star(r) = y$. For $t \in \{0, \dots, n\}$, suppose that $g^\star(s^t) = y$ and $g^\star(s^{t+1}) \neq z$. If $g^\star(s^{t+1}) \notin \{y, z\}$, then $s^t \neq s^{t+1}$ so that $t + 1 \in N \setminus K_x(p)$ and hence $y \succ_{t+1}^{s^{t+1}} g^\star(s^{t+1})$. Therefore, individual $t + 1$ could manipulate $g^\star \equiv g|_{C_x^\star}$ at s^{t+1} via $s^t(t+1)$, a contradiction. Therefore, $g^\star(s^t) \in \{y, z\}$ for all $t \in \{0, \dots, n\}$.

Suppose that $g^\star(s^t) = y$ for all $t \in \{0, \dots, n\}$, so that in particular $g^\star(s) \equiv g^\star(s^n) = y$. Let profile $w \in C_x$ be such that $g|_{C_x} = z$, which necessarily exists since $\#g(C_x) \geq 3$. Let $\{s^t\}_{t=0}^n$ be the modified standard sequence from w to s in which the preferences of the individuals in $K_x(p)$ are changed before those of the individuals in $N \setminus K_x(p)$. For $t \in \{0, \dots, \kappa - 1\}$, $\#N_{s^t}(x, z) \geq \#N_w(x, z) > \frac{n}{2}$ for all $z \in g(C_x) \setminus \{x\}$, since individual preference ordering $w(t)$ is being replaced with $s(t)$ for which $s_1(t) = x$. For $t \in \{\kappa, \dots, n\}$, $\#N_{s^t}(x, z) \geq \#K_x(s^t) \geq \#K_x(s) = \kappa - 1 > \frac{n}{2}$ for all $z \in g(C_x) \setminus \{x\}$. Therefore, $\{s^t\}_{t=0}^n \subset C_x$.

By definition, $g|_{C_x}(s^0) \equiv g|_{C_x}(w) = z$. For $t \in \{0, \dots, \kappa - 2\}$, suppose that $g|_{C_x}(s^t) = z$ and $g|_{C_x}(s^{t+1}) = y$. Since individual $t + 1 \in K_x(p)$, $z \succ_{t+1}^{s^{t+1}} y$, so that individual $t + 1$ can manipulate $g|_{C_x}$ at profile s^{t+1} via $s^t(t + 1)$, a contradiction.

Therefore, $g|_{C_x}(s^t) = z$ for $t \in \{0, \ldots, \kappa - 1\}$.

Similarly, for $t \in \{\kappa, \ldots, n\}$, suppose that $g|_{C_x}(s^t) = z$ and $g|_{C_x}(s^{t+1}) = y$. Since $z \succ_{t+1}^{s^{t+1}} y$, individual $t+1$ can manipulate $g|_{C_x}$ at profile s^{t+1} via $s^t(t+1)$, a contradiction. Therefore, $g|_{C_x}(s^t) = z$ for $t \in \{\kappa, \ldots, n\}$. In particular, $g|_{C_x}(s^n) \equiv g|_{C_x}(s) \equiv g^\star(s) = z$. Therefore, for an arbitrary $z \in g(C_x) \setminus \{x, y\}$, $z \in g^\star(C_x^\star)$, which implies that $g^\star(C_x^\star) = g(C_x)$.

The restriction $g^\star$ can be used to induce a social choice rule $\hat{g} : L(X)^{n-\kappa+1} \to g(C_x)$ the set of individuals in $N \setminus K_x(p)$ in the natural way. The main result of Aswal, Chatterji, and Sen (2003) establishes the Gibbard–Satterthwaite theorem (1973, 1975) on the domain $L(X_x)^m$ for arbitrary finite m; in particular, we may take $m = n-\kappa+1$. Therefore, if $\hat{g}$ is strategy–proof – a property inherited from $g|_{C_x}$ through $g^\star$ by the nature of restrictions – it must be dictatorial with dictator $h \in N \setminus K_x(p)$.

Choose an arbitrary $z \in g(C_x) \setminus \{x, y\}$ and let profile v be such that $v(h) = (z \succ x \succ \ldots)$ and $v(i) = (x \succ \cdots \succ z)$ for all $i \in N \setminus \{h\}$. Note that for all $y \in X \setminus \{x\}$, $\#N_v(x, y) \geq \#K_x(v) = n - 1 > \frac{n}{2}$ since by assumption $n > 2$, so that $v \in C_x$. Define profile v' such that $v'(i) = r(i)$ for all $i \in K_x(p)$ and $v'(i) = v(i)$ for all $i \in N \setminus K_x(p)$. Note that for all $y \in X \setminus \{x\}$, $\#N_{v'}(x, y) \geq \#K_x(v') \geq \#K_x(p) > \frac{n}{2}$, so that $v' \in C_x^\star$. Then $g|_{C_x}(v') \equiv g^\star(v') = v'_1(h) = z$, by definition.

Let $\{v^t\}_{t=0}^n$ is the standard sequence from v' to v. For all $t \in \{0, \ldots, n\}$, $\#N_{v^t}(x, y) \geq \#K_x(v^t) \geq \#K_x(v) > \frac{n}{2}$ for all $y \in X \setminus \{x\}$, so that $\{v^t\}_{t=0}^n \subset C_x$. Note that if for any $t \in \{0, \ldots, n\}$, if $g|_{C_x}(v^t) \notin \{x, z\}$, then individual h can manipulate $g|_{C_x}$ at v^t via any preference ordering $s(t)$ such that $s_1(t) = x$, since $x \succ_h^{v^t} g|_{C_x}(v^t)$ and by the induction hypothesis, $\#K_x(s) = \kappa$ implies that $g|_{C_x}(s)$ for any profile $s \in C_x$ such

that $s(i) = v^t(i)$ for all $i \in N \setminus \{h\}$ and $s_1(t) = x$. Therefore, $g|_{C_x}(v^t) \in \{x, z\}$ for all $t \in \{0, \ldots, n\}$.

From above, $g|_{C_x}(v^0) \equiv g|_{C_x}(v') = z$. For $t \in \{0, \ldots, n\}$, suppose that $g|_{C_x}(v^t) = z$ and $g|_{C_x}(v^{t+1}) = x$. Then $v^t \neq v^{t+1}$, which implies that $t+1 \in K_x(p)$, so that $g|_{C_x}(v^{t+1}) \succ^{v^t}_{t+1} g|_{C_x}(v^t)$ and individual $t+1$ can manipulate $g|_{C_x}$ at v^t via $v^{t+1}(t+1)$, a contradiction. Therefore, for $t \in \{0, \ldots, n\}$, $g|_{C_x}(v^t) = z$ implies $g|_{C_x}(v^{t+1}) = z$. In particular $g|_{C_x}(v^n) \equiv g|_{C_x}(v) = z$.

Since v was any profile for which $v(i) = (x \succ \cdots \succ z)$ for all $i \in N \setminus (K_x(p) \cup \{h\})$, standard sequence arguments imply that $g|_{C_x}(v'') = z$ for all $v'' \in C_x$ such that $v''_1(h) = z$. Similarly, since $z \in g(C_x) \setminus \{x, y\}$ was chosen arbitrarily, standard sequence arguments imply that $g|_{C_x}(v''') = y$ for all $v''' \in C_x$ such that $v'''_1(h) = y$, for all $y \in g(C_x) \setminus \{x\}$.

It therefore remains to show that $g|_{C_x}(v''') = x$ for all $v''' \in C_x$ such that $v'''_1(h) = x$. Note that the arguments of the previous paragraph do not suffice in this case, as the position of the Condorcet winner x within individual preference orderings cannot be chosen arbitrarily while remaining in the domain C_x.

For some $z \in C_x \setminus \{x\}$, let $v \in C_x$ be such that $v(h) = (x \succ z \succ \ldots)$ and $y \succ^v_i z$ for all $y \in g(C_x) \setminus \{z\}$ and all $i \in N \setminus \{h\}$. By previous arguments, $g|_{C_x}(v) \in \{x, z\}$; otherwise h could manipulate $g|_{C_x}$ at v via $v' \in C_x$, where $v'(i) = v(i)$ for all $i \in N \setminus \{h\}$ and $v'_1(h) = z$.

Let profile v' be such that $v'(h) = v(h)$ and for all $i \neq h$, $v'(i)$ is formed from $v(i)$ by promoting alternative x to the top of the preference ordering. Note that $\#N_{v'}(x, y) \geq \#K_x(v') = n$, so that $v' \in C_x$ and by the induction hypothesis, $g|_{C_x}(v') = x$.

Let $\{v'^t\}_{t=0}^n$ be the standard sequence from v to v', and note that for all $t \in \{0,\ldots,n\}$, $\#N_{v'^t}(x,y) \geq \#N_v(x,y) > \frac{n}{2}$ for all $y \in X \setminus \{x\}$, since by assumption $v \in C_x$. By definition, $g|_{C_x}(v'^0) \equiv g|_{C_x}(v) = z$. For $t \in \{0,\ldots,n\}$, Suppose that $g|_{C_x}(v'^t) = z$ and $g|_{C_x}(v'^{t+1}) \neq z$. If $g|_{C_x}(v'^{t+1}) \neq x$ or $t+1 \in N \setminus \{h\}$, then individual $t+1$ can manipulate $g|_{C_x}$ at v'^{t+1} via $v'^t(t+1)$, since $z \succ_{t+1}^{v'^{t+1}} g|_{C_x}(v'^{t+1})$. If $g|_{C_x}(v'^{t+1}) = x$ and $t+1 = h$, then individual $t+1$ can manipulate $g|_{C_x}$ at v'^t via $v'^{t+1}(t+1)$, since $x \succ_{t+1}^{v'^{t+1}} z$. Therefore, $g|_{C_x}(v'^t) = z$ implies that $g|_{C_x}(v'^{t+1}) = z$, so that by induction $g|_{C_x}(v') \equiv g|_{C_x}(v'^n) = z$, a contradiction. Therefore, it must be that $g|_{C_x}(v) = x$.

Next, let $v'' \in C_x$ be any profile for which $v_1''(h) = x$. Let $y \in g(C_x) \setminus \{x\}$ denote the maximal element in $g(C_x) \setminus \{x\}$ under $\succ_h^{v''}$; that is, y is the second–most preferred alternative in $g(C_x)$ at profile v'' by individual h, which necessarily exists since $\#g(C_x) \geq 3$.

Consider a profile v' such that $v''(h) = v'(h)$ and for all $i \in N \setminus \{h\}$ $v'(i)$ is formed from $v''(i)$ by demoting alternative y to the bottom of individual i's preference ordering at v'. Since $\#N_{v''}(x,z) > \frac{n}{2}$ implies $\#N_{v'}(x,z) > \frac{n}{2}$ for all $z \in X \setminus \{x\}$, $v' \in C_x$. From the arguments of previous paragraphs, $g|_{C_x}(v') = x$.

Suppose that $\#N_{v''}(x,y) = \alpha$, and let $\{v''^t\}_{t=0}^n$ be the modified standard sequence from v' to v'' in which the preferences of individual h are changed first, followed by the remaining $\alpha - 1$ individuals in $N_{v''}(x,y)$ are changed first. Note that $\#N_{v''}(x,z) > \frac{n}{2}$ implies $\#N_{v''^t}(x,z) > \frac{n}{2}$ for all $z \in X \setminus \{x\}$, so that $\{v''^t\}_{t=0}^n \subset C_x$.

From the arguments above, $g|_{C_x}(v''^t) \in \{x,y\}$ for all $t \in \{0,\ldots,n\}$, and $g|_{C_x}(v') = g|_{C_x}(v''^0) = x$. For $t \in \{0,\ldots,\alpha\}$, $g|_{C_x}(v''^t) = x$; otherwise $g|_{C_x}(v''^t) = x$ and

$g|_{C_x}(v''^{t+1}) = y$, so that individual $t+1$ can manipulate $g|_{C_x}$ at v''^{t+1} via $v''^{t}(t+1)$, since $x \succ_{t+1}^{v''^{t+1}} y$.

Therefore, $g|_{C_x}(v''^{t}) = x$ for $t \in \{a+1, \ldots, n\}$; otherwise $g|_{C_x}(v''^{t}) = x$ and $g|_{C_x}(v''^{t+1}) = y$, so that individual $t+1$ can manipulate $g|_{C_x}$ at v''^{t} via $v''^{t+1}(t)$, since $y \succ_{t+1}^{v''^{t}} x$. Therefore, $g|_{C_x}(v'') \equiv g|_{C_x}(v''^{n}) = x$, so that $g|_{C_x}(v'') = x$ for any $v'' \in C_x$ such that $v_1''(h) = x$.

Collecting the arguments of the previous paragraphs, if $g^\star$ is dictatorial with dictator h, then $g|_{C_x}$ itself has individual h as the dictator over the alternatives in $g(C_x)$.

Step 1.3. From the previous two steps, if $g|_{C_x}$ is non–dictatorial then $g|_{C_x}(p) = x$ for any $p \in C_x$ for which $\#K_x(p) > \frac{n}{2}$. Assuming that $g|_{C_x}$ is non–dictatorial, it remains to prove that $g|_{C_x}(p) = x$ if $\#N_p(x,y) > \frac{n}{2}$ for all $y \in X \setminus \{x\}$; that is, at all $p \in C_x$.

Suppose that $g|_{C_x}$ is non–dictatorial and there exists a $p \in C_x$ such that $g|_{C_x}(p) = y$ for $y \in g(C_x) \setminus \{x\}$. Let profile q be such that $q_1(i) = x$ for all $i \in N_p(x,y)$ and $q(i) = p(i)$ for all $i \in N_p(y,x)$, and note that $\#N_p(x,z) > \frac{n}{2}$ implies that $\#N_q(x,z) > \frac{n}{2}$ for all $z \in X \setminus \{x\}$, so that $q \in C_x$.

Let $\{q^t\}_{t=0}^n$ be the modified standard sequence from p to q in which the preferences of the individuals in $N_p(x,y)$ are changed first. Note that $\#N_p(x,z) > \frac{n}{2}$ implies that $\#N_{q^t}(x,z) > \frac{n}{2}$ for all $z \in X \setminus \{x\}$, so that $\{q^t\}_{t=0}^n \subset C_x$. By assumption, $g|_{C_x}(p) \equiv g|_{C_x}(q^0) = y$. For $t \in \{0, \ldots, \#N_p(x,y) - 1\}$, suppose that $g|_{C_x}(q^t) \neq x$ but $g|_{C_x}(q^{t+1}) = x$. Then individual $t+1$ can manipulate $g|_{C_x}$ at q^t via $q^{t+1}(t+1)$, since $x \succ_{t+1}^{q^{t+1}} g|_{C_x}(q^t)$, a contradiction. Therefore, $g|_{C_x}(q^t) \neq x$ for $t \in \{0, \ldots, \#N_p(x,y)\}$.

By Step 2 above, $g|_{C_x}(q^t) = x$ for $t \in \{\#N_p(x,y)+1, \ldots, n\}$, since $\#K_x(q^t) = \#N_p(x,y) > \frac{n}{2}$. However, by the construction of q, $q^{\#N_p(x,y)+1} = q^{\#N_p(x,y)}$, a contradiction. Therefore, $g|_{C_x}(p) = x$ for all profiles $p \in C_x$, a contradiction since $\#g(C_x) \geq 3$. Therefore, if $g|_{C_x}$ is strategy–proof must be dictatorial with respect to the alternatives in C_x.

Part 2: Suppose that $g|_{C_x}$ is strategy–proof and $x \notin g(C_x)$.

Step 2.1. Define the subdomain U_x as follows: for all $p \in U_x$ and all $i \in N$, $p_1(i) = x$, $p_k(i) \in g(C_x)$ for $k \in \{2, \ldots, \#g(C_x)+1\}$, and $(p_{\#g(C_x)\#+2}(i) \succ p_{\#g(C_x)\#+3}(i) \succ \ldots)$ is some fixed ordering of the alternatives in $X \setminus (g(C_x) \cup x)$ for all $i \in N$. Note that for all $p \in U_x$, $\#K_p(x) = n > \frac{n}{2}$, so that $U_x \subset C_x$ and therefore $g(U_x) \subseteq g(C_x)$.

For $y \in g(C_x)$, let $p \in C_x$ be such that $g|_{C_x}(p) = y$ and let $u \in U_x$ be such that $p_2(i) = y$ for all $i \in N$. Let $\{u^t\}_{t=0}^n$ denote the standard sequence from p to u, and note that $\#N_{u^t}(x,z) \geq \#N_p(x,z) > \frac{n}{2}$ for all $z \in X \setminus \{x\}$, so that $\{u^t\}_{t=0}^n \subset C_x$.

By construction, $g|_{C_x}(u^0) \equiv g|_{C_x}(p) = y$. For $t \in \{\#N_p(x,y)+1, \ldots, n-1\}$, suppose that $g|_{C_x}(u^t) = y$ and $g|_{C_x}(u^{t+1}) \in X \setminus \{x,y\}$. Then individual $t+1$ can manipulate $g|_{C_x}$ at u^{t+1} via $u^t(t+1)$, since $y \succ_{t+1}^{u^{t+1}} g|_{C_x}(u^{t+1})$, a contradiction. Therefore, $g|_{C_x}(u^t) = y$ for $t \in \{\#N_p(x,y)+1, \ldots, n\}$ and hence $g(C_x) \subseteq g(U_x)$, so that $g(U_x) = g(C_x)$.

Note that within U_x, preferences with respect to the elements of $g(U_x)$ are unrestricted and $\#g(U_x) \geq 3$. The Gibbard–Satterthwaite theorem (1973, 1975) therefore implies that $g|_{C_x}$ must be dictatorial over U_x with respect to the set of alternatives $g(C_x)$. Without loss of generality suppose that individual 1 is the dictator.

Step 2.2. Let $p \in U_x$ be such that $g|_{C_x}(p) = y$, and $p^{(1)} \in C_x$ be any profile such that $p^{(1)}(i) = p(i)$ for all $i\{2, \ldots, n\}$. Let $y \in g(C_x)$ be maximal in $g(C_x)$ under $p^{(1)}(1)$. If $g|_{C_x}(p^{(1)}) \neq y$, then individual 1 can manipulate $g|_{C_s}$ at $p^{(1)}$ via $p(1)$, since $y \succ_1^{p^{(1)}} g|_{C_x}(p^{(1)})$, a contradiction. Therefore, $g|_{C_x}$ is dictatorial with respect to $g(C_x)$ with individual 1 as the dictator at all such profiles $p^{(1)}$.

Next consider any profile $p^{(2)} \in C_x$ such that $p^{(2)}(i) = p^{(1)}(i)$ for all $i\{1, 3, \ldots, n\}$. If $g|_{C_x}(p^{(2)}) \succ_2^{p^{(2)}} y$, then individual 2 can manipulate $g|_{C_x}$ at $p^{(1)}$ via $p^{(2)}(2)$, a contradiction. Further, if $y \succ_2^{p^{(2)}} g|_{C_x}(p^{(2)})$, then individual 2 can manipulate $g|_{C_x}$ at $p^{(2)}$ via $p^{(1)}(2)$, a contradiction. Therefore, $g|_{C_x}(p^{(2)}) = y$ and hence $g|_{C_x}$ is dictatorial with respect to $g(C_x)$ with individual 1 as the dictator at all such profiles $p^{(2)}$. Proceeding inductively, $g|_{C_x}$ is dictatorial with respect to $g(C_x)$ with individual 1 as the dictator at all such profiles $p^{(k)}$, for $k \in \{1, \ldots, n\}$.

The entire domain C_x can be reconstructed from such an sequence of profiles $p^{(k)}$. Note that from any $q \in C_x$ there exists a profile $u \in U_x$ at which each individual's relative ordering of the alternatives in X_x is identical to their relative ordering at q. Let $\{q^t\}_{t=0}^n$ be the standard sequence from u to q, and note that $\#N_{q^t}(x, z) \geq \#N_q(x, z) > \frac{n}{2}$ by definition, so that $\{q^t\}_{t=0}^n \subset C_x$. Furthermore, q^t corresponds to a profile of the type $p^{(t)}$ from above, so that individual 1 is the dictator at all q^t. In particular, individual 1 is the dictator at an arbitrary profile $q \in C_x$. $\square$

Appendix B

Appendix to Chapter 3

B.1 Commitment Mechanism Usage In ATE

Recall Equation 3.26,

$$c_T > \left(\frac{1-\delta+\beta\delta}{1-\delta}\right) c_M, \tag{B.1}$$

which embodies joint conditions on c_T, c_M, β, and δ under which the sophisticated agent will employ the commitment mechanism of membership termination using an alternating termination equilibrium. Note that

$$\frac{d}{d\delta}\left[\frac{1-\delta+\beta\delta}{1-\delta}\right] = \frac{\beta}{(1-\delta)^2} > 0, \tag{B.2}$$

which implies that that Equation 3.26 is satisfied at some $\bar{\delta}$, it is satisfied at all $\delta > \bar{\delta}$ holding all other parameters fixed, since $c_T, c_M < 0$ and $\beta > 0$. Similarly,

$$\frac{d}{d\beta}\left[\frac{1-\delta+\beta\delta}{1-\delta}\right] = \frac{\delta}{1-\delta} > 0, \tag{B.3}$$

which implies that that Equation 3.26 is satisfied at some $\bar{\beta}$, it is satisfied at all $\beta > \bar{\beta}$ holding all other parameters fixed, since $c_T, c_M < 0$ and $\delta > 0$.

B.2 Proof of Theorem 3.3.1

Proof. The result follows readily from Billingsley (1968) and Dobb's Martingale Convergence Theorem.

Following Feldman (1991), from the perspective of the learning–naïve agent the set of possible complete descriptions of $\hat{\beta}$, the time–invariant value of the quasi–hyperbolic discount factor entering into the utility calculations of past– and future–period selves, can be represented as a separable metric space Θ with Borel σ–field $B(\Theta)$. Uncertain as to the value of her true quasi–hyperbolic discount factor, $\beta \in \Theta$, the agent is endowed with prior beliefs μ_1 on $(\Theta, B(\Theta))$[1] and an induced probability P_{μ_1} on an infinite horizon belief–outcome space. Let $\{\mu_t\}_{t=1}^{\infty}$ denote the sequence of posterior beliefs, and note that since f has positive on $\left(\frac{-c_E}{\beta\delta}, \frac{-c_E}{\delta}\right)$, there is sufficient variance in the stochastic benefit process for the learning–naïve agent to realize both periods of maintenance and periods of exercise. Noting that the sequence of beliefs are a martingale with respect to the probability P_{μ_1}, it follows from the Martingale Convergence Theorem (Doob, 1953) that the Bayesian beliefs converge almost surely (with respect to P_{μ_1}) to some limit belief μ_∞.

Given that $\{\mu_t\}_{t=1}^{\infty}$ converges almost surely to μ_∞, suppose that μ_∞ does not coincide with the degenerate distribution that puts unit probability on $\hat{\beta} = \beta$; that is, μ_∞ does not coincide with the true distribution of β. If the learning–naïve agent

has not yet terminated her membership, belief updating proceeds according to

$$\begin{aligned}\mu_t(x) &\equiv \mu_{t-1}(x|d_{t-1}=E,b_{t-1}) \\ &= \begin{cases} 0, & \text{if } 0 \leq x \leq \dfrac{-c_E}{\delta b_{t-1}} \\ \dfrac{\mu_{t-1}(x)}{\displaystyle\int_{-c_E/\delta b_{t-1}}^{1} \mu_{t-1}(y)dy}, & \text{if } \dfrac{-c_E}{\delta b_{t-1}} < x \leq 1 \end{cases}\end{aligned} \tag{B.4}$$

and

$$\begin{aligned}\mu_t(x) &\equiv \mu_{t-1}(x|d_{t-1}=M,b_{t-1}) \\ &= \begin{cases} \dfrac{\mu_{t-1}(x)}{\displaystyle\int_{0}^{-c_E/\delta b_{t-1}} \mu_{t-1}(y)dy}, & \text{if } 0 \leq x \leq \dfrac{-c_E}{\delta b_{t-1}} \\ 0, & \text{if } \dfrac{-c_E}{\delta b_{t-1}} < x \leq 1, \end{cases}\end{aligned} \tag{B.5}$$

which are not stationary processes for generic draws from the distribution f on B. Therefore, if μ_∞ does not coincide with the true distribution of β, it follows that the agent previously terminated her membership in some (finite) period τ.

If μ_∞ coincides with the true distribution of β, then by the convergence it follows that for every $\epsilon > 0$, there exists a $\tau \in \{1, 2, 3, \dots\}$ such that for every $t \geq \tau$,

$$\left|\mathbb{E}_{\mu_t}[\hat{\beta}] - \mathbb{E}_{\mu_\infty}[\hat{\beta}]\right| < \epsilon. \tag{B.6}$$

Choosing $\epsilon \in (0, \bar{\beta} - \beta)$, there exists a (finite) τ such that for all $t \geq \tau$, $\mathbb{E}_{\mu_t}[\hat{\beta}] < \bar{\beta}$. Therefore, the agent will terminate her membership in period τ. □

B.3 Mathematica Simulation

Mathematica Simulation of the Partially–Naïve Agent of Subsection 3.3.5.

```
Clear[beta, delta, cE, bbar, benchmark, t, b, M, r, count];

beta = 1/4;

delta = 1/2;

cE = -25;

bbar = 60;

benchmark = (3*(53*Sqrt[106]-557))/(5*(71*Sqrt[106]-785));


count = 0;

r = {};


(*Perform 1,000 simulation runs*)

While[count < 1000,


 (*Reset delay count*)

 t = 0;


 (*Reset prior beliefs*)

 mu =.;

 mu[x_] := Piecewise[{{1/100, 0 <= x <= 98/99}, {4901/50, 98/99 < x <= 1}}];


 (*Simulate single decision problem*)

 While[NIntegrate[x*mu[x], {x, 0, 1}] > benchmark,


   (*Generate realization of benefit to exercising*)

   Clear[b];
```

```
  b = 10*(5 - Log[1 - RandomReal[{0, 1}]]);


  (*Choose current-period optimal action*)
  Clear[M];
  M = If[cE + (beta*delta*b) > 0, 0, 1];


  (*Update beliefs based on last period's action*)
  temp =.;
  temp[x_] :=
   Simplify[
    M*Piecewise[{{mu[x]/
          Integrate[mu[y], {y, 0, Min[-cE/(delta*b), 1]}],
         0 <= x <= Min[-cE/(delta*b), 1]}, {0,
         Max[-cE/(delta*b), 0] < x <= 1}}] + (1 - M)*
      Piecewise[{{0, 0 <= x <= Min[-cE/(delta*b), 1]}, {
         mu[x]/Integrate[mu[y], {y, Max[-cE/(delta*b), 0], 1}],
         Max[-cE/(delta*b), 0] < x <= 1}}]];


  mu =.;
  mu[x_] = temp[x];
  t++;
  ]


(*Save number of periods prior to termination*)
Clear[tempR];
tempR = Append[r, t];
r = tempR;
count++;
```

```
]
```

Appendix C

Appendix to Chapter 4

C.1 Proof of Lemma 4.3.4

Proof. Let Γ be a biclique network and $\succ \in \mathbf{P}$ be given. By Gale & Shapley (1962), there exists $\mu \in \mathbf{U}(\mathcal{M}, \mathcal{W}, \succ)$. Since Γ is biclique, $(m, \mu(m)) \in \Gamma$ for all $m \in \mathcal{M}$, and naturally $(w, \mu(w)) \in \Gamma$ for all $w \in \mathcal{W}$; hence, μ respects Γ. Moreover, since μ is a stable matching, it is individually rational and has no blocking pairs. In particular, this implies that μ has no local blocking pairs on Γ. Therefore, $\mu \in \mathbf{U}(\mathcal{M}, \mathcal{W}, \succ)$ implies that $\mu \in \mathbf{U}_T^\Gamma(\mathcal{M}, \mathcal{W}, \succ) \subseteq \mathbf{U}_I^\Gamma(\mathcal{M}, \mathcal{W}, \succ)$.

Suppose next that $\mu \in \mathbf{U}_T^\Gamma(\mathcal{M}, \mathcal{W}, \succ)$ but $\mu \notin \mathbf{U}(\mathcal{M}, \mathcal{W}, \succ)$. Since μ is technologically Γ–stable, it must be individually rational; hence, if $\mu \notin \mathbf{U}(\mathcal{M}, \mathcal{W}, \succ)$ then it must be the case that μ is blocked by some blocking pair $(m, w) \in \mathcal{M} \times \mathcal{W}$. Having assumed that Γ is biclique, $(m, w) \in \Gamma$ so that (m, w) is in fact a local blocking pair for μ on Γ, a contradiction. Therefore, $\mu \in \mathbf{U}_T^\Gamma(\mathcal{M}, \mathcal{W}, \succ)$ implies that $\mu \in \mathbf{U}(\mathcal{M}, \mathcal{W}, \succ)$.

Following the logic of the previous paragraph, $\mu \in \mathbf{U}_I^\Gamma(\mathcal{M}, \mathcal{W}, \succ)$ implies that $\mu \in \mathbf{U}(\mathcal{M}, \mathcal{W}, \succ)$. Therefore, when Γ is biclique

$$\mathbf{U}(\mathcal{M}, \mathcal{W}, \succ) = \mathbf{U}_T^\Gamma(\mathcal{M}, \mathcal{W}, \succ) = \mathbf{U}_I^\Gamma(\mathcal{M}, \mathcal{W}, \succ) \tag{C.1}$$

for any preference profile $\succ \in \mathbf{P}$. □

C.2 Proof of Proposition 4.3.5

Proof. Let $\mathcal{M}$, $\mathcal{W}$, and $\succ \in \mathbf{P}$ be given such that there exist $m \in \mathcal{M}$ and $w \in \mathcal{W}$ with $w \succ_m m$ and $m \succ_w w$. Define a network Γ on $\mathcal{M} \times \mathcal{W}$ as follows: $(m', w') \in \Gamma$ if and only if $m' \succ_{m'} w'$ or $w' \succ_{w'} m'$. That is, Γ consists only of links between man–woman pairs for whom one of the parties finds the other unacceptable.

By Gale & Shapley (1962), there exists $\mu \in \mathbf{U}(\mathcal{M}, \mathcal{W}, \succ)$. To achieve a contradiction, suppose that for every $\mu \in \mathbf{U}(\mathcal{M}, \mathcal{W}, \succ)$, $\mu(m') = m'$ for all $m' \in \mathcal{M}$ and $\mu(w') = w'$ for all $w' \in \mathcal{W}$. By assumption, $w \succ_m m$ and $m \succ_w w$, so that $(m, w) \in \mathcal{M} \times \mathcal{W}$ could block μ, a contradiction. Therefore, there exists a matching $\mu \in \mathbf{U}(\mathcal{M}, \mathcal{W}, \succ)$ at which at least one man is matched. Define $(\hat{m}, \hat{w}) \in \mathcal{M} \times \mathcal{W}$ such that $\mu(\hat{m}) = \hat{w}$.

Note that by construction μ does not respect Γ: the individual rationality of μ implies that $\hat{w} \succ_{\hat{m}} \hat{m}$ and $\hat{m} \succ_{\hat{w}} \hat{w}$, so that $(\hat{m}, \hat{w}) \notin \Gamma$. Hence, $\mu \in \mathbf{U}(\mathcal{M}, \mathcal{W}, \succ)$ but $\mu \notin \mathbf{U}_T^{\Gamma}(\mathcal{M}, \mathcal{W}, \succ)$, so that $\mathbf{U}(\mathcal{M}, \mathcal{W}, \succ) \not\subseteq \mathbf{U}_T^{\Gamma}(\mathcal{M}, \mathcal{W}, \succ)$.

By Theorem 4.3.8, there exists $\nu \in \mathbf{U}_T^{\Gamma}(\mathcal{M}, \mathcal{W}, \succ)$. Moreover, $\nu(m') = m'$ for all $m' \in \mathcal{M}$, since μ is individually rational and for all $w' \in \mathcal{W}$, $(m', w') \in \Gamma$ implies that $m' \succ_{m'} w'$ or $w' \succ_{w'} m'$, by construction. Consequently, $\nu(w') = w'$ for all $w' \in \mathcal{W}$.

Note that ν is not stable on the marriage problem $\langle \mathcal{M}, \mathcal{W}, \succ \rangle$; by assumption $w \succ_m m \equiv \nu(m)$ and $m \succ_w w = \nu(w)$, so that $(m, w) \in \mathcal{M} \times \mathcal{W}$ blocks ν. Hence, $\nu \in \mathbf{U}_T^{\Gamma}(\mathcal{M}, \mathcal{W}, \succ)$ but $\nu \notin \mathbf{U}(\mathcal{M}, \mathcal{W}, \succ)$, so that $\mathbf{U}_T^{\Gamma}(\mathcal{M}, \mathcal{W}, \succ) \not\subseteq \mathbf{U}(\mathcal{M}, \mathcal{W}, \succ)$, as desired. □

C.3 Proof of Proposition 4.3.7

Proof. Let $\mathcal{M}$, $\mathcal{W}$, and $\Gamma \subsetneq \mathcal{M} \times \mathcal{W}$ be given. Since Γ is not biclique, there exists $(\hat{m}, \hat{w}) \in \mathcal{M} \times \mathcal{W}$ such that $(\hat{m}, \hat{w}) \notin \Gamma$.

Consider any preference profile $\succ \in \mathbf{P}$ such that

(i) $\hat{w} \succ_{\hat{m}} w$ for all $w \in \mathcal{W} \setminus \{\hat{w}\}$, and;

(ii) $\hat{m} \succ_{\hat{w}} m$ for all $m \in \mathcal{M} \setminus \{\hat{m}\}$.

By Gale & Shapley (1962), there exists $\mu \in \mathbf{U}(\mathcal{M}, \mathcal{W}, \succ)$. If $\mu(\hat{m}) \neq \hat{w}$, then $(\hat{m}, \hat{w})$ can block μ, since by (i) and (ii) above, $\hat{w} \succ_{\hat{m}} \mu(\hat{m})$ and $\hat{m} \succ_{\hat{w}} \mu(\hat{w})$. Therefore, $\mu(\hat{m}) = \hat{w}$, so that μ does not respect Γ. Hence, $\mu \in \mathbf{U}(\mathcal{M}, \mathcal{W}, \succ)$ but $\mu \notin \mathbf{U}_T^{\Gamma}(\mathcal{M}, \mathcal{W}, \succ)$, so that $\mathbf{U}(\mathcal{M}, \mathcal{W}, \succ) \not\subseteq \mathbf{U}_T^{\Gamma}(\mathcal{M}, \mathcal{W}, \succ)$.

By Theorem 4.3.8, there exists $\nu \in \mathbf{U}_T^{\Gamma}(\mathcal{M}, \mathcal{W}, \succ)$. Since ν respects Γ, $\nu(\hat{m}) \neq \hat{w}$. However, by (i) and (ii) above, $(\hat{m}, \hat{w})$ is a blocking pair for ν on the marriage problem $\langle \mathcal{M}, \mathcal{W}, \succ \rangle$ since $\hat{w} \succ_{\hat{m}} \nu(\hat{m})$ and $\hat{m} \succ_{\hat{w}} \nu(\hat{w})$. Hence, $\nu \in \mathbf{U}_T^{\Gamma}(\mathcal{M}, \mathcal{W}, \succ)$ but $\nu \notin \mathbf{U}(\mathcal{M}, \mathcal{W}, \succ)$, so that $\mathbf{U}_T^{\Gamma}(\mathcal{M}, \mathcal{W}, \succ) \not\subseteq \mathbf{U}(\mathcal{M}, \mathcal{W}, \succ)$, as desired. □

C.4 Proof of Theorem 4.3.8

Proof. Given a marriage network $\langle \mathcal{M}, \mathcal{W}, \succ, \Gamma \rangle$, the Men–Proposing Network–Respecting Deferred Acceptance Algorithm (NDAA) proceeds as detailed below. The Women–Proposing NDAA is defined analogously by reversing the roles of the men and the women.

Let $\langle \mathcal{M}, \mathcal{W}, \succ, \Gamma \rangle$ be a given marriage network. For each agent $m \in \mathcal{M}$ construct a new preference profile $\succ'_m$ from $\succ_m$ and Γ as follows:

(i) For all $w \in \mathcal{W}$ such that $m \succ_m w$, let $m \succ'_m w$;

(ii) For all $w \in \mathcal{W}$ such that $(m, w) \notin \Gamma$, let $m \succ'_m w$;

(iii) For all $w, \tilde{w} \in \mathcal{W}$ such that $w \succ_m \tilde{w} \succ_m m$, $(m, w) \in \Gamma$, and $(m, \tilde{w}) \in \Gamma$, let $w \succ'_m \tilde{w} \succ'_m m$.

Note that preferences over unacceptable women at $\succ'_m$ have not been uniquely specified; as in the standard matching literature, stability results on marriage networks are robust to arbitrary re–orderings of the relative preference ranking of unacceptable agents.

Applying the analogous operations, construct preferences $\succ'_w$ from $\succ_w$ for all $w \in \mathcal{W}$. Let $\succ'$ denote the newly constructed preference profile, and let μ be the stable matching selected by the Men–Proposing Deferred Acceptance Algorithm (Gale & Shapley, 1962) on the marriage problem $\langle \mathcal{M}, \mathcal{W}, \succ' \rangle$.

By the stability of μ on $\langle \mathcal{M}, \mathcal{W}, \succ' \rangle$, μ is individually rational with respect to $\succ'$. That is, for all $m \in \mathcal{M}$, if $\mu(m) \in \mathcal{W}$ then $\mu(m) \succ'_m m$. In particular, this implies that if $\mu(m) \in \mathcal{W}$, then $(m, \mu(m)) \in \Gamma$ by the construction of $\succ'$. Applying the same argument to the set of women, the matching μ respects Γ. Moreover, since $w \succ'_m m$ implies that $w \succ_m m$ by the construction of $\succ'$, individual rationality of μ with respect to $\succ'$ implies individual rationality with respect to $\succ$.

It remains to show that μ has no local blocking pairs on Γ at preference profile $\succ$. Suppose that $(m, w) \in \mathcal{M} \times \mathcal{W}$ forms a local blocking pair; that is, $w \succ_m \mu(m)$, $m \succ_w \mu(w)$ and $(m, w) \in \Gamma$. By the construction of $\succ'$, this implies that $w \succ'_m \mu(m)$ and $m \succ'_w \mu(w)$. Therefore, (m, w) is a blocking pair for μ under $\succ'$, a contradiction to the stability of μ.

Therefore $\mu \in \mathbf{U}_T^{\Gamma}(\mathcal{M}, \mathcal{W}, \succ)$, so that there exists a technologically Γ–stable matching on the marriage network $\langle \mathcal{M}, \mathcal{W}, \succ, \Gamma \rangle$. □

C.5 Proof of Proposition 4.3.10

Proof. For a fixed marriage network $\langle \mathcal{M}, \mathcal{W}, \succ, \Gamma \rangle$, suppose there exists a matching $\mu \in \mathbf{U}_T^{\Gamma}(\mathcal{M}, \mathcal{W}, \succ)$ that is Pareto dominated by a matching ν that respects Γ. That is, $\nu \succsim_i \mu$ for all $i \in \mathcal{N}$, and $\nu \succ_j \mu$ for some $j \in \mathcal{N}$.

Since $(j, \nu(j)) \in \Gamma$, if $\nu \succ_{\nu(j)} \mu$, then $(j, \nu(j))$ is a local blocking pair for μ, a contradiction. Therefore, $\nu \sim_{\nu(j)} \mu$. Since agents have strict preferences over matching partners, this implies that $j = \mu(\nu(j))$, which contradicts the assumption that $\nu \succ_j \mu$.

Therefore, the set of technologically Γ–stable matchings is Pareto dominant in the set of matchings that respect Γ. □

C.6 Proof of Proposition 4.3.11

Proof. Let $\succ$, $\mathcal{M}$, and $\mathcal{W}$ be such that there exists $m \in \mathcal{M}$ and $w \in \mathcal{W}$ for whom $w \succ_m m$ and $m \succ_w w$. Define $\Gamma \subset \mathcal{M} \times \mathcal{W}$ to be such that $(m', w) \notin \Gamma$ for all $m' \in \mathcal{M}$ and $(m, w') \notin \Gamma$ for all $w' \in \mathcal{W}$. That is, Γ is such that man m and woman w have no acquaintances in Γ.

Let $\mu \in U_I^\Gamma(\mathcal{M}, \mathcal{W}, \succ)$ be such that $\mu(m) = m$ and $\mu(w) = w$. Note that such an informationally Γ–stable matching exists by applying Corollary 4.3.9 to the marriage network $\langle \mathcal{M}', \mathcal{W}', \succ|_{\mathcal{M}' \times \mathcal{W}'}, \Gamma|_{\mathcal{M}' \times \mathcal{W}'} \rangle$, where $\mathcal{M}' \equiv \mathcal{M} \setminus \{m\}$ and $\mathcal{W}' \equiv \mathcal{W} \setminus \{w\}$. Since man m and woman w have no acquaintances in Γ, they cannot be members of any local blocking pair of μ on Γ; moreover, by construction μ has no local blocking pairs that do not contain either man m or woman w.

Consider the matching ν such that $\nu(m) = w$ and $\nu(i) = \mu(i)$ for all $i \in \mathcal{N} \setminus \{m, w\}$, and note that $\nu \in U_I^\Gamma(\mathcal{M}, \mathcal{W}, \succ)$ since μ had no local blocking pairs, $w \succ_m m$, and $m \succ_w w$. Note moreover that $\nu \succsim_i \mu$ for all $i \in \mathcal{N}$, and $\nu \succ_i \mu$ for $i \in \{m, w\}$. Hence, $\mu \in U_I^\Gamma(\mathcal{M}, \mathcal{W}, \succ)$ is Pareto dominated by $\nu \in U_I^\Gamma(\mathcal{M}, \mathcal{W}, \succ)$. □

C.7 Proof of Theorem 4.3.15

Proof. Let μ be the stable matching selected by the Men–Proposing Network–Respecting Deferred Acceptance Algorithm on the marriage network $\langle \mathcal{M}, \mathcal{W}, \succ, \Gamma \rangle$. By the definition of technological achievability, showing that μ is the $\mathcal{M}_T$–optimal stable matching is equivalent to showing that for each $m \in \mathcal{M}$, $\mu(m)$ is maximal with respect to $\succ_m$ in $\mathbf{A}_T^\Gamma(m; \mathcal{M}, \mathcal{W}, \succ)$. As in Gale & Shapley (1962), this can be shown by demonstrating that no man is ever rejected by a technologically achievable woman in the Men–Proposing Network–Respecting Deferred Acceptance Algorithm; since it is incentive compatible for men to make proposals in the order of their preferences, this implies that each man is matched with his most–preferred technologically achievable woman.

The proof is by induction. Suppose that no man has yet been rejected by a woman who is technologically achievable to him when woman $w \in \mathcal{W}$ rejects man $m \in \mathcal{M}$. Note that this implies that $(m, w) \in \Gamma$, as man m cannot propose to women with whom he is unacquainted. If $w \succ_w m$, then w is not technologically achievable for man m, and the next round of proposals begins with no man having ever been rejected by a woman who is technologically achievable to him.

If $m \succ_w w$ but woman w rejects man m, she must have received a proposal from some man $m' \in \mathcal{M}$ such that $m' \succ_w m$. By the definition of the algorithm, $(m', w) \in \Gamma$ and man m' prefers woman w to any woman except those who have already rejected him and those with whom he is

unacquainted, who, by the inductive argument, are necessarily unachievable to him.

Consider a hypothetical matching ν that respects Γ such that $\nu(m) = w$ and everyone else is matched to a technologically achievable partner (or left unmatched, if no technologically achievable partners remain). By the above argument, $w \succ_{m'} \nu(m')$, $m' \succ_w \nu(w)$, and $(m', w) \in \Gamma$. Therefore, (m', w) is a local blocking pair for ν in Γ.

Note that the set of matchings from which ν was drawn includes the set of all technologically network–stable matchings at which man m is matched to woman w. Therefore, there is no technologically Γ–stable matching that matches m and w, so that woman w is technologically unachievable for man m, and the next round of proposals begins with no man having ever been rejected by a woman who is technologically achievable to him.

It therefore follows that $\mu(m)$ is maximal in $\mathbf{A}_T^\Gamma(m; \mathcal{M}, \mathcal{W}, \succ)$ with respect to $\succ_m$, so that the matching produced by the Men–Proposing Network–Respecting Deferred Acceptance Algorithm is indeed $\mathcal{M}_T$–optimal. An analogous proof shows that the matching produced by the Women–Proposing Network–Respecting Deferred Acceptance Algorithm is indeed $\mathcal{W}_T$–optimal. □

C.8 Proof of Theorem 4.3.18

Proof. Let $\mu, \mu' \in \mathbf{U}_T^\Gamma(\mathcal{M}, \mathcal{W}, \succ)$ and suppose that $\mu \succsim_{\mathcal{M}} \mu'$ and there exists $w \in \mathcal{W}$ such that $\mu \succ_w \mu'$.

Since μ and μ' are individually rational, agents' preferences are strict, and $\mu \succ_w \mu'$, there exists $m \in \mathcal{M}$ such that $\mu(w) = m$; in particular, this implies that $(m, w) \in \Gamma$. Moreover, $\mu \succ_w \mu'$ implies that $\mu'(m) \neq w$, since otherwise $\mu(w) = \mu'(w)$ and therefore $\mu \sim_w \mu'$.

Therefore, there exist $(m, w) \in \mathcal{M} \times \mathcal{W}$ such that $(m, w) \in \Gamma$, $w \succ_m \mu'(m)$, and $m \succ_w \mu'(w)$, so that (m, w) forms a local blocking pair for μ' on Γ, a contradiction.

Therefore, $\mu \succsim_{\mathcal{M}} \mu'$ implies that $\mu' \succsim_{\mathcal{W}} \mu$. A symmetric argument proves the reverse implication.

□

C.9 Proof of Theorem 4.3.20

Proof. Let $\langle \mathcal{M}, \mathcal{W}, \succ, \Gamma \rangle$ be a marriage network with $\mu, \nu \in \mathbf{U}_T^{\Gamma}(\mathcal{M}, \mathcal{W}, \succ)$, and define $\lambda = (\mu \wedge \nu)$ to be the meet of μ and ν.

For $w \in \mathcal{W}$, suppose that $\lambda(w) = m$ for some $m \in \mathcal{M}$ and that, without loss of generality, $\mu(w) = m$. This implies that $m \succsim_w \nu(w)$ and $(m, w) \in \Gamma$, by the definition of the meet. If $\lambda(m) \neq w$, then the definition of the meet further implies that $\lambda(m) = \nu(m) \neq w$ and hence $w \succ_m \nu(m)$. Therefore, $(m, w) \in \mathcal{M} \times \mathcal{W}$ is a local blocking pair for ν, a contradiction. Hence, $\lambda(w) = m$ implies $\lambda(m) = w$ for all $(m, w) \in \mathcal{M} \times \mathcal{W}$.

To see the reverse implication, define

$$\begin{aligned} \mathcal{W}' &= \{w \in \mathcal{W} \,:\, \lambda(w) \in \mathcal{M}\} \\ &= \{w \in \mathcal{W} \,:\, \mu(w) \in \mathcal{M} \text{ or } \nu(w) \in \mathcal{M}\}, \end{aligned} \tag{C.2}$$

where the last equality follows from the fact that μ and ν are individually rational and each woman w gets her maximal match from $\{\mu(w), \nu(w)\}$. By our above result, we have that

$$\begin{aligned} \lambda(\mathcal{W}') &\subseteq \{m \in \mathcal{M} \,:\, \lambda(m) \in \mathcal{W}\} \\ &= \{m \in \mathcal{M} \,:\, \mu(m) \in \mathcal{W} \text{ and } \nu(m) \in \mathcal{W}\} \equiv \mathcal{M}', \end{aligned} \tag{C.3}$$

where the last equality follows from the fact that μ and ν are individually rational and each man m gets his minimal match from $\{\mu(m), \nu(m)\}$. Note that $|\mathcal{M}'| = |\mu(\mathcal{M}')|$ and $|\lambda(\mathcal{W}')| = |\mathcal{W}'|$ since μ and λ are one–to–one matchings. Moreover, by construction $|\mathcal{W}'| \geq |\mu(\mathcal{M}')|$, since every matched partner of each $m \in \mathcal{M}'$ under μ is included in $\mathcal{W}'$ by definition. Note that $\lambda(\mathcal{W}') \subseteq \mathcal{M}'$ implies that $|\lambda(\mathcal{W}')| \leq |\mathcal{M}'|$, so that the above system of inequalities implies that $|\lambda(\mathcal{W}')| = |\mathcal{M}'|$ and hence $\lambda(\mathcal{W}') = \mathcal{M}'$.

Hence, if $m \in \mathcal{M}'$, then $\lambda(m) = w$ for some $w \in \mathcal{W}'$, so $\lambda(w) = m$ by the above reasoning. If $m \notin \mathcal{M}'$, $\lambda(m) = m$. That is, if $\lambda(m) = w$ for some $(m, w) \in \mathcal{M} \times \mathcal{W}$, then $\lambda(w) = m$. Combining both implications shows that λ is a well–defined matching, with $\lambda(m) = w$ if and only if $\lambda(w) = m$, for all $(m, w) \in \mathcal{M} \times \mathcal{W}$.

It is straightforward to see that λ is individually rational and respects Γ whenever μ and ν are technologically Γ–stable. It therefore remains to show that λ does not have any local blocking pairs. To achieve a contradiction, suppose that $(m, w) \in \mathcal{M} \times \mathcal{W}$ is a local blocking pair of λ; that is, $(m, w) \in \Gamma$, $w \succ_m \lambda(m)$ and $m \succ_w \lambda(w)$. Without loss of generality, suppose that $\lambda(m) = \nu(m) \neq w$. Then by the construction of λ and the transitivity of preferences, $w \succ_m \nu(m)$ and $m \succ_w \lambda(w) \succsim_w \nu(w)$, so that (m, w) are a local blocking pair for ν, a contradiction.

Therefore, $\lambda = (\mu \wedge \nu) \in \mathbf{U}_T^{\Gamma}(\mathcal{M}, \mathcal{W}, \succ)$; that is, λ is a well–defined matching that is individually rational, respects Γ, and has no local blocking pairs. Analogous arguments show that $\lambda = (\mu \vee \nu) \in \mathbf{U}_T^{\Gamma}(\mathcal{M}, \mathcal{W}, \succ)$. □

C.10 Proof of Theorem 4.4.2

Proof. Given a network formation game $\langle \mathcal{M}, \mathcal{W}, \{u_i\}_{i \in \mathcal{N}} \rangle$, let $\mu \in \mathbf{U}_T^{\Gamma}(\mathcal{M}, \mathcal{W}, \succ)$, where the preference relation $\succ_i$ is represented by the utility function u_i for all $i \in \mathcal{N}$. Define pure–strategy profile s be such that $s_i(j) = 1$ if and only if $\mu(i) = j$.

Holding the strategy profiles of players $\mathcal{N} \setminus \{i\}$ fixed, consider the best–response strategy of player i. Since $s_j(i) = 0$ for all $j \neq \mu(i)$, any strategy that sets $s_i(j) = 1$ for $j \neq \mu(i)$ is strictly dominated by $s_i(j) = 0$; proposing a link to such a player j will incur player i the propositional cost ϵ but will not change the set of player with whom player i can potentially match, since the link (i, j) will not be formed. Similarly, the bounds on c insure that the matching–utility of all players in any stable matching exceeds the cost of link formation, so that setting $s_i(\mu(i)) = 1$ dominates $s_i(\mu(i)) = 0$. Hence, pure strategy s_i is a best response to s_{-i}, so that pure–strategy profile s constitutes a pairwise–Nash equilibrium.

The reverse implication follows immediately from Lemma 4.4.3 and definition of stability. □

C.11 Proof of Lemma 4.4.3

Proof. Given a network formation game $\langle \mathcal{M}, \mathcal{W}, \{u_i\}_{i \in \mathcal{N}} \rangle$, suppose that $s^\star$ is a pure–strategy pairwise–Nash equilibrium of the network formation game preceding the selection of a technologically network–stable matching. To achieve a contradiction, suppose that $\mu, \nu \in \mathbf{U}_T^{\Gamma(s^\star)}(\mathcal{M}, \mathcal{W}, \succ)$, where $\Gamma(s^\star)$ is the network resulting (deterministically) from strategy profile $s^\star$ and the preference relation $\succ_i$ is represented by the utility function u_i for all $i \in \mathcal{N}$.

If $\mu \neq \nu$, then there exists $i \in \mathcal{N}$ such that $\mu(i) \neq \nu(i)$. Since $\succ_i$ is complete and asymmetric, without loss of generality it must be the case that $\mu(i) \succ_i \nu(i)$. Then $\mathbb{E}[u_i(\Gamma(s^\star_{-i}, \tilde{s}_i)] > \mathbb{E}[u_i(\Gamma(s^\star))]$, where $\tilde{s}_i(j) = s_i(j)$ for $j \neq \nu(i)$ and $\tilde{s}_i(\nu(i)) = 0$; strict inequality is guaranteed since $\nu \in \mathbf{U}_T^{\Gamma(s^\star)}(\mathcal{M}, \mathcal{W}, \succ)$ implies that $s_i^\star(\nu(i)) = 1$. Since this contradicts the assumption that $s^\star$ is a pure–strategy pairwise–Nash equilibrium, it must be that $\mu = \nu$ and hence $\mathbf{U}_T^{\Gamma(s^\star)}(\mathcal{M}, \mathcal{W}, \succ) = \{\mu\}$. Moreover, repeated application of the above reasoning shows that $\mu(i) = j$ if and only if $s_i^\star(j) = 1$. □